University of California
Santa Barbara

Multimodal Analytics for Healthcare

A dissertation submitted in partial satisfaction
of the requirements for the degree

Doctor of Philosophy
in
Electrical and Computer Engineering

by

Carlos Torres

Committee in charge:

Professor B. S. Manjunath, Chair
Jeffrey C. Fried, M.D.
Scott D. Hammond, M.D.
Professor Joao Pedro Hespanha

January 2018

The Dissertation of Carlos Torres is approved.

Jeffrey C. Fried, M.D.

Scott D. Hammond, M.D.

Professor Joao Pedro Hespanha

Professor B. S. Manjunath, Committee Chair

November 2017

Multimodal Analytics for Healthcare

by

Carlos Torres

"Thank you... for gracing my life with your lovely presence, for adding the sweet measure of your soul to my existence."

To my Bothers, Sisters, Mom (RIP), and my Family: Elya, Eryn, Baby, and dogs.

I love you all!

Acknowledgements

I want to thank my sisters and brothers for their never ending love, support, and help. I owe you guys everything. I want to thank my wife and friend, Eryn for her support, patience, and for all the nights she did not sleep reading my drafts. Also, I want to thank my friends and most closest collaborators: Amir M. Rahimi, Archith J. Bency, Aruna R. Jammalamadaka, Carter De Leo, Christopher Wheat, Diana L. Delibatov, Dmitry L. Fedorov, and Victor E. Fragoso for their continuous support and words of confidence and motivation. They deserve a round of applause for listening to my ranting and venting over the past years.

Special thanks to all my undergraduate advisors and professors from San Jose State University Daryl K. Eggers, Herbert Siber (RIP 2017), and Karen Singmaster; my undergraduate advisors at the Georgia Institute of Technology: Charles C. Kemp and Gary May; and graduate professors and advisors at the University of California Santa Barbara: B. S. Manjunath, Christian Villasenor, Joao P. Hespanha, Scott D. Hammond, Jeffrey C. Fried, and Lawrence Rabiner. Without their patience I would not have started on this path or seen it through.

Curriculum Vitæ
Carlos Torres

Education

September 2017	**Doctor's of Philosophy (PhD).** Department of Electrical and Computer Engineering, University of California Santa Barbara.
June 2012	**Master's of Science (MS).** Signal Processing & Machine Intelligence. Department of Electrical and Computer Engineering. University of California Santa Barbara. Santa Barbara, California.
June 2009	**Bachelor's of Science (BS).** Dual Degree in Electrical Engineering and Bio-engineering with Concentrations in Digital and Analog Systems Design from the Electrical Engineering Department and College of Engineering. Minors in Chemistry, Bio-Chemistry, Physics, and Math, San Jose State University. San Jose, California.

Fields of Study

Computer vision and machine learning methods for healthcare

Experience

2011 – 2017	Graduate Research Assistant, Vision Research Lab (VRL), University of California, Santa Barbara, CA.
2016 – Present	Procore Technologies. Data Sciences, Computer Vision, and Machine Learning Researcher. Carpinteria, CA.
2015 – 2016	Carpe Data. Chief Data Scientist. Santa Barbara, CA.
2015	Caugnate. Computer Vision Software Developer. Goleta, CA.
2007 – 2009	Research Assistant, Egger's Laboratory. San Jose State University. San Jose, CA.
2008	Research Assistant, Healthcare Robotics Laboratory. Georgia Institute of Technology. Atlanta, GA.
2007	Research Assistant, Hewlett-Packard Labs. Palo Alto, CA.

Publications

Carlos Torres, Jeffrey C. Fried, and B. S. Manjunath. HEAL: Healthcare Event and Activity Analysis and Logging. *Trans. IEEE / EMBS Journal of Translational Engineering in Health and Medicine (JTEHM).* March 2018. – **To Appear**

Carlos Torres, Jeffrey C. Fried, Kennenth Rose, and B. S. Manjunath. MASH: Multimodal Multiview Motion Analysis and Summarization for Healthcare. *Trans. IEEE Multimedia. Emerging Areas: Healthcare.* January 2018. – **To Appear**

Carlos Torres, Archith J. Bency, Jeffrey C. Fried, and B. S. Manjunath. RAM: Role Representation and Identification from combined Appearance and Activity Maps. *Proc. IEEE/ACM Int'l. Conference on Distributed Smart Cameras (ICDSC).* Stanford, California. September 7-9, 2017.

Carlos Torres, Kenneth Rose, Jeffrey C. Fried, and B. S. Manjunath. Deep Eye-CU (DECU): Summarization of Patient Motion in the ICU. *Proc. European Conference on Computer Vision (ECCV).* Amsterdam, The Netherlands. October 8-16, 2016.

Carlos Torres, Victor Fragoso, Scott D. Hammond, Jeffrey C. Fried, and B. S. Manjunath.Eye-CU: Sleep Pose Classification for Healthcare using Multimodal Multiview Data.*Proc. IEEE Winter Conference on Applications of Computer Vision (WACV).* Lake Placid, NY, USA. March 7-9, 2016.

Carlos Torres, Scott D. Hammond, Jeffrey C. Fried, and B. S. Manjunath. Sleep Pose Recognition in an ICU From Multimodal Data and Environmental Feedback. *Proc. Springer Int'l. Conference on Computer Vision Systems (ICVS).* Copenhagen, Denmark. July 6-9. Vol. 9163. Springer, 2015.

Phillip J. Calabretta, Mitchell C. Chancellor, **Carlos Torres**, Gary R. Abel, Jr., Clayton Niehaus, Nathan J. Birtwhistle, Nada M. Khouderchah, Genet H. Zemede, and Daryl K. Silica as a Matrix for Encapsulating Proteins: Surface Effects on Protein Structure Assessed by Circular Dichroism Spectroscopy. *Journal of functional biomaterials.* Vol.3. No. 3 . 514-527, 2012.

B. Menaa, **C. Torres**, M. Herrero, V. Rives, A.R.W. Gilbert, and D.K. Eggers. Protein Adsorption onto Organically Modified Silica Glass Leads to a Different Structure than Sol-Gel Encapsulation. *Biophysical journal.* Vol.95. No. 8. L51-L53, 2008.

Abstract

Multimodal Analytics for Healthcare

by

Carlos Torres

The ailing healthcare system demands effective autonomous solutions to improve service and provide individualize care, while reducing the burden on the scarce healthcare workforce. Most of these solutions require a multidisciplinary approach that combines healthcare with computational abilities. The work presented in this thesis introduces a multimodal multiview network along with methods and solutions that leverage inexpensive visual sensors and computers to monitor healthcare. One of the most prominent outcomes of this work includes enabling the medical analysis of ICU conditions such as sleep disorders, decubitus ulcerations, and hospital acquired infections, which are preventable and negatively affect patients' health population. The problems tackled include patient pose classification, pose motion analysis and summarization, role representation and identification, and activity and event logging in natural hospital settings. These problems are addressed via a non-intrusive non-disruptive multimodal multiview sensor network (Medical Internet-of-Things). The multimodal data is combined with coupled-optimization to estimate source weights and accurately classify patient poses. Pose patterns such as pose transitions are represented using deep convolutional features and pose duration is modelled via segments. The proposed techniques serve to differentiate between poses and pseudo-poses (transitory poses) and create effective motion summaries. The role representation is tackled using novel appearance and semantic interaction maps to assign generic labels to individuals (doctor, nurse, visitor, etc) without using identifiable information (e.g., facetracking or badges), which is prohibited in health-

care applications. Finally, activity and event analysis is tackled using a new contextual aspect frames where aspect bases and weights are learned and then used to reconstruct activities. The objective of this thesis is to enable the development, evaluation, and optimization of individualized therapies, standards-of-care, room infrastructural designs, and clinical workflows and procedures.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

We can only see a short distance ahead, but we can see plenty there that needs to be done. -Alan Turing

Much needed technological overhaul of the healthcare system in the U.S. includes changing its focus from disease to patient centric, and by providing the means to individualize care. For example, clinical evidence suggests that body poses of patients on beds are correlated with patient recovery rates and response to therapies. There are two major, connected limitations in the design and dissemination of pose-related healthcare protocols. First, manual analysis is the most effective, but it requires staff to track and record patient poses. Second, automated monitoring systems that can remove the burden from human observers have been unreliable in natural scenarios, which can be partially occluded, poorly illuminated, and continuously changing (e.g., moving equipment). These issues are addressed in this thesis by introducing methods for robust monitoring of Intensive Care Unit (ICU) scenarios using a non-intrusive and non-disruptive multimodal and multiview sensor network.

The diagram in Figure 1.1 shows the organization of this dissertation. The two main parts are the static and dynamic analysis of healthcare environments.

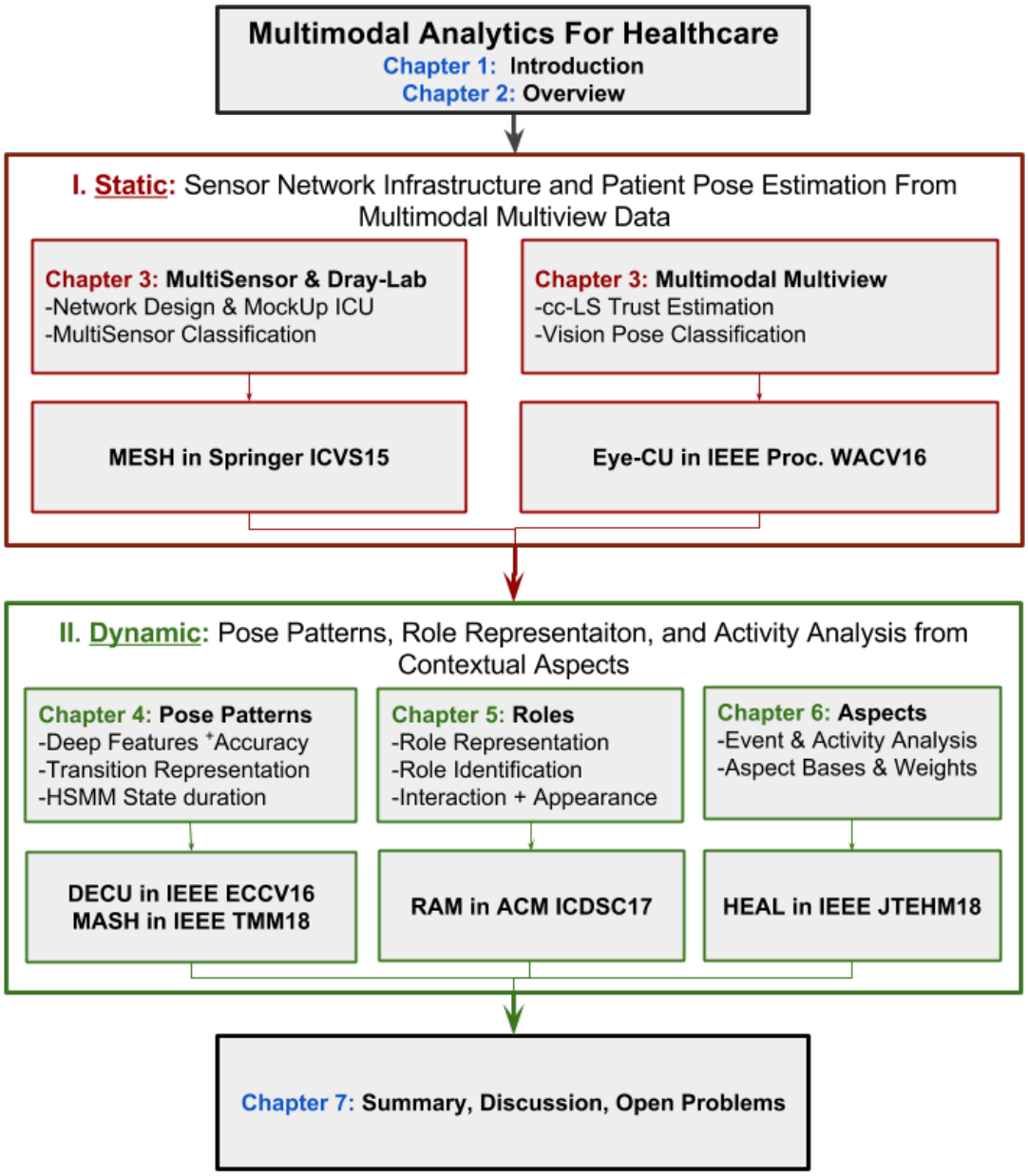


Figure 1.1: Thesis organization.
This thesis has two main parts: static (red) and dynamic (green) analysis.

The two main themes are: static (red) and dynamic (green) analysis. The static portion is covered in Chapter 3. The dynamic portion is covered in Chapters 4, 5, and 6. A detailed summary of this thesis is provided in Chapter 7 along with a comprehensive

list of open problems and future work.

1.1 Motivation

Evidence-based medicine implores scientists and clinicians to collect and process all available data in a specific healthcare setting. New methods for non-disruptive monitoring and analysis of patient-on-bed body configurations, e.g., sleep-pose patterns, add objective metrics for predicting and evaluating health related scenarios. clinical examples where the identification of body poses of patients is correlated to medical conditions such as sleep apnea, gastroesophageal reflux disease (GERD), decubitus ulcerations, pulmonary disorders, and back pain. For example, sleep positions can amplify or attenuate the symptoms of sleep apnea (i.e., snoring) – where the obstructions of the airway are only present or are greatest in supine (laying on back) positions; and the symptoms of GERD are reduced by laying on the side. Body positioning is important in acute lung injury and prone positioning has been shown to improve outcomes in adult respiratory distress syndrome; and back and spine problems often worsen when patients lay on their backs or stomachs, so lateral positioning of patients is recommended by medical experts. During pregnancies, physicians recommend that mothers-to-be lay on their sides to improve fetal blood flow. Although the standard of care in most ICU patients who are relatively immobile, e.g., on a mechanical ventilator, is to turn these patients every two hours to prevent decubitus ulcers (i.e., bed sores), which is a protocol with very low compliance rates. Clinical studies correlate sleep positions (among other causes) to various effects on health and quality of sleep of ICU patients and non-patients. In particular, these studies clearly indicate that studying sleep poses in natural sleep scenarios helps to evaluate sleep and to improve diagnosis and optimize treatment of sleep disorders.

In particular, Harvard Medical School reported in August 2016 that monitoring ICUs

can save up to \$15 billion by saving $\$20,000$ in each of the $750,000$ ICU beds in the U.S. by reducing the effect of preventable ICU-related conditions such as poor quality of sleep and decubitus ulcers (DUs) [53]. For instance, ICUs in the U.S. receive about five million patients per year, each with an average stay of 9.3 days and with a mortality rate that ranges from 10 to 30% depending on health conditions. MASH sample applications focus on developing solutions to help understand and address sleep analysis and incidence of DUs. These applications are selected due to their pervasive nature in medical ICUs and the opportunity to improve the quality of care provided to patients. For example, sleep hygiene is correlated to shorter hospital stays, increased recovery rates, and decreased mortality rates.

1.2 Challenges and Contributions

The main challenges include: existing hospital infrastructure which requires the network to be sanitizable (i.e., enclosed) and to work with existing standards-of-care and prohibits devices from having exposed wires, real-world hospital (i.e., natural) scene conditions, subtle and erratic patient motion, obfuscated identities and roles, privacy and HIPAA regulations, and complex activities and events. The details of specific challenges are described as follows:

1. Multimodal Sensor Network and Static Pose Analysis. The first challenge described in Chapter 3, refers to the need to address healthcare deficiencies while avoiding disrupting existing standards of care. A mock-up ICU room serves to design, deploy, evaluate, iterate, and refine a data collection system (i.e., multimodal multivew sensor network), methods, and algorithms for healthcare. More importantly, these can be designed, deployed, and tested without disrupting patients or their care. The first task: pose classification is tackled using a multisensor system that uses a single RGB-Depth

camera and a high resolution pressure array. The pressure requires direct contact with the patients, hence medical protocols demand the array to be properly sterilized before and after being used. The pressure array's cost and complexity make it logistically impossible to use in healthcare settings. The system must be markerless, sterilizable, simple and easy to deploy, and must work with existing hospital infrastructure.

Although multimodal multiview patient pose classification helps dealing with illumination variations and partial occlusions, the challenges arises when two or more sources conflict on candidate poses. The problem changes from pose classification to modality (or view) trust estimation for pose classification tasks.

The contributions from the work presented in Chapter 3 focus on a multimodal sensor network for pose classification. The first section of this thesis focuses on the design, development, and deployment of a non-intrusive, non-obtrusive sensor network. The first element of the network involves sensor performance evaluation for specific activities that may enable the analysis of sleep deprivation, decubitus ulcerations, and hospital acquired infections. The studies use a multi-sensor network, which evolved into a purely observation multimodal multiview network. The contributions of this first stage include:

- Multi-Sensor Patient Pose Classification with Environmental Feedback The multi-sensor (rgb, depth, and pressure) single view network is a reliable network for patient pose classification evaluated in a simulated scenarios over a vast range of scene conditions with variable non-uniform illumination and partial occlusions such as pillows and blankets.

- Multimodal Multiview Pose Classification and Modality Trust The multimodal multiview network is a markerless, purely visual network that avoids contact with patients. The method uses coupled-optimization to estimate source (modality and view) contributions of each sensor for accurate pose classification under for each of

the evaluated scene conditions.

2. Pose Pattern Analysis and Summarization. The pose classification challenges are expanded to incorporate pose pattern analysis, which is also detailed in Chapter 3. The challenges include effective pose feature and pose transition representation (i.e., pseudo-poses); and differentiating between pose histories (slow sequences over long periods of time: seconds) and pose transitions (fast sequences over short periods of time: hours). These challenges are addressed using deep features to better represent static poses, multimodal multiview keyframe extraction based on frame dissimilarities, and time-series analysis using time slices to represent transitions and segments to model pose duration and differentiate between pose histories and pose transitions. The contributions from this work includes the dynamic (i.e., temporal) analysis of people (patients, visitors and staff) motion within the ICU context. The first part of dynamic analysis encompasses the patient rate and range of motion. The characterization and typification of patient pose motion. This objective required effective pose sequence and pose transition representation and quantification over long and short periods of time (slow or fast motion). Specifically the contributions are:

- Pose Transition Representation and Analysis Improved pose representation is achieved with deep convolutional neural network features extracted from Google's Inception architecture. The transition are represented using keyframes, which are extracted using a proposed algorithm that deals spans the multiple modalities and views, which are available the sensor network used to collect pose transition data.

- Pose Pattern Analysis and Summarization Sleep patterns range from seconds to hours. A time-segment based method derived from Markov Modelling is implemented to more flexible model state duration and distinguish between slow and fast pose patterns.

3. Healthcare Events and Activity Analysis. Activity and event analysis in healthcare is paramount to the optimization of therapies, services, and patient care. A critical element is the identification of individuals in and their activities in the ICU room. However, privacy protection restrictions prohibit the use of personal identifiable information such as faces, badges, or identification cards. Therefore, the new concept of roles is introduced. Roles are learned from each individuals as a combination of appearance (clothing color and textures) and interaction maps (orientation and relative distances over time). The interactions are represented as time-evolving histograms, which encode relative orientation and distances to tagged elements in the ICU such as medical equipment, computers, patients, entrance, and sink.

Roles serve to distinguish common and specific activities that each person performs in the ICU. However, more information is necessary to clearly and accurately identify activities. For instance, hand washing and hand sanitizing can be easily confused, unless other aspects such as sink vs sanitizing gel, short duration vs long duration, and towel vs no towel differences are identified and highlighted (i.e., the aspects) However, most of the aspects are not clearly identify and many more objects are present (or absent). The challenge is two fold: on the one hand, the aspects need to be identified. On the second had, the aspects need to be properly weighted based on their activity reconstruction abilities. Basically, what aspects and how prevalent each aspect is in the observed set of training activities.

The contributions of this work are encapsulated by the second stage of the dynamic analysis portion of this dissertation. The work focuses on more generalized human motion analysis and uses interaction and activity maps to characterize and infer people roles. The applications of the findings include methods for activity and event analysis and logging, which can help understand deficiencies and optimize workflows and sanitation procedures. These contributions include:

- Role representation and identification A new method to represent roles by combining appearance and interaction dynamics is introduced. The method enables identifying events in healthcare while preserving the privacy of individuals in the ICU room.

- Contextual Aspects for Activity and Event Analysis Activities are commonly identified by body dynamics, but omit using contextual information. One more contribution of this thesis is the introduction of contextual aspects for activity representation and classification. Aspects are the attributes such as roles, objects, duration, etc. Aspects bases and aspect weights are the identify aspects and their respective contributions in activity reconstruction.

1.3 Organization

This dissertation is organized as follows:

- Chapter 2 provides an overview of the problems in healthcare tackled by this work. In addition, it provides an overview of technical related work and motivates the need for the work discussed in this dissertation.

- Chapter 3 studies the classification of sleep poses in the natural healthcare scenarios. It describes the design of the multimodal sensor network (MESH) and its evolution. Also, this Chapter introduces the methods for static sleep pose classification (EYE-CU) for actors in a mock-up ICU room and for patients in two medical ICU rooms.

- Chapter 4 incorporates temporal patterns to the analysis of patient sleep poses. In particular it introduces a framework to model pose duration and differentiate between poses (long periods) and pseudo-poses or transitory poses. The proposed

Hidden Semi-Markov Model (HSMM) pose pattern analysis methods are derived form Hidden Markov Model (HMM) techniques to more flexible model state duration. The analysis of poses includes sequence of poses (history) and pose transitions (transition angular range and direction of rotation).

- Chapter 5 extends the study beyond patient sleep poses. It looks at the various individuals that visit the ICU room and deals with privacy protection stipulations from HIPAA. This Chapter introduces the concept of healthcare roles to assign visitors one of various labels, without ever having to directly them. A new framework for learning role representations from appearance vectors and activity semantic maps (RAM) is proposed in this Chapter. In particular, the semantic maps framework reinforces role representation over time. For instance, role definitions are enhanced by a person's interactions and locations visited while in the ICU. The semantic maps are very helpful when roles are required but ICU conditions require all visitors to wear isolation scrubs (i.e., everyone looks the same).

- Chapter 6 introduces activity contextual aspects and their applications for event analysis and logging (HEAL). In addition to formalizing the concept of contextual aspects for activity analysis, this chapter uses the network from Chapter 3 and incorporates the findings from Chapter 4 and Chapter 5 to analyze and log activities and events in a natural ICU setting.

- Chapter 7 re-iterates the main concepts of this dissertation, summarizes its contributions, and provides an overview of potential future directions that include behavior analysis for healthcare study and its applications.

Chapter 2

Overview of Human Pose and Behavior Analysis in Healthcare

The oldest, shortest words "yes" and "no" are those which require the most thought. -Pythagoras of Samos

2.1 Introduction

The tenets of evidence-based medicine implore clinicians and researchers to collect and process all available data in a specific healthcare setting. For example, clinical anecdotal information indicates that poses can negatively or positively affect patients health status. Ability to detect and monitor patient poses enables the objective design and evaluation of positioning therapies. Dynamically, patient poses can be analyzed to estimate and quantify motion (rate and range) and how measured changes and patterns can be used to understand common ICU disorder such as sleep deprivation, sleep hygiene, and the

incidence and prevention of decubitus ulcerations (i.e., bed sores). This information can be used to typify and evaluate patient behavior. In addition to pose pattern analysis, dynamic analysis includes analyzing non-patient motion and behavior in the form of activities and events and correlate these with patient health.

New methods for non-disruptive monitoring and analysis of patient sleep poses, patterns, and quality add objective metrics for predicting and evaluating health-related scenarios. There are clear clinical examples where patient poses are correlated to medical conditions. For example, the symptoms of gastroesophageal reflux disease (GERD) are reduced by laying on the side [33]. Body positioning is important in acute lung injury and prone positioning has been shown to improve outcomes in adult respiratory distress syndrome [23]. Prone and supine positions worsen back and spine problems, so lateral positioning is recommended by medical experts [20]. Physicians recommend that pregnant women lay on their sides to improve fetal blood flow [49].

2.2 Correlations between Pose Patterns, Human Behavior, and Patient Health

There is a clear healthcare correlation between patient pose patterns and human behavior and patient health status. Computer vision methods for pose classification and monitoring are used in [35,40,59] to detect and classify sleep poses but are limited to ideal scenes. In both approaches, the staging needed for observation affects the measurements. In order to overcome these issues, the solutions proposed in this thesis begin with the use of three non-invasive, independent sensor modalities: RGB, depth, and pressure and evolved into a purely visual (no pressure needed) multiview multimodal sensor network and analysis algorithms.

Existing techniques are able to estimate human poses in ideal scenes using these modalities independently, but they fail in challenging ones. In [87] the authors present a generative approach that uses deformable parts model (DPM), commonly used in RGB images. Unfortunately, the DPM method requires images with relatively uniform illumination and with only minor self-occlusions. The discriminative approach from [66] uses depth images and is robust to illumination changes. However, this method requires clean depth segmentation and contrast, and it fails under occlusions. Neither of these methods works in unconstrained ICU scenarios.

Decubitus Ulcerations. Harvard Medical School reported in August 2016 that monitoring ICUs can save up to $15 billion by saving $20,000 in each of the 750,000 ICU beds in the U.S. by reducing the effect of preventable ICU-related conditions such as poor quality of sleep and decubitus ulcers (DUs) also known as bed sores [53]. For instance, ICUs in the U.S. receive about five million patients per year, each with an average stay of 9.3 days and with a mortality rate that ranges from 10 to 30% depending on health conditions, about 80% of those patients will be at risk of developing bed sores. DUs are preventable, soft tissue wounds that appear on bony areas of the body and are caused by continuous decubitus positions. The standard of care for immobile ICU patients is to rotate them every two hours to prevent decubitus ulcers, but this is rarely accomplished or effective [68]. There is little understanding about the set of poses and pose duration that cause or prevent DUs. The methods introduced in this thesis enable the inception of required clinical studies that analyze pose duration, rotation frequency and range, and the duration of weight/pressure off-loading, as well as serving as the non-obtrusive measuring tool to collect and analyze pose patterns.

Sleep Disorders. In particular, sleep positions affect the symptoms of conditions such as sleep apnea – where airway obstructions are greatest in supine positions [64]. There are two major approaches for the study of sleep. One approach uses bio-status data to monitor a patient's metabolic state during sleep [34, 38, 59]. The polysomnogram is the standard equipment used in these studies. Its motion-restricting probes connect to the patient's head, face, and respiratory system, monitoring brain activity, rapid-eye-motion (REM) signals, and levels of oxygen and carbon-dioxide in the blood. The second approach is based on the identification of sleep patterns using non-intrusive equipment and human observers [26, 40]. The findings in [5, 29, 83] correlate body positions to various effects on health and quality of sleep of ICU patients. The authors state that identification of sleep poses in natural scenarios helps to evaluate sleep and to improve diagnosis and treatment of sleep disorders. Current physiological systems use machines that physically connect to the patients, making them disruptive and intrusive. Purely observational systems use images and pressure arrays to estimate poses but have been unable to handle natural scenes – indoor ICU scenes with variable illumination and occlusions such as blankets and pillows.

Hospital Acquired Infections. Consider the issue of Hospital Acquired Infections (HAIs) by touch in the ICU. For consistency, assume that events can have one of four qualifiers: clean, contamination, transmission, or unclean. The labels depend on the sequence of underlying actions and the detection of hand sanitation actions performed by people after entering and before exiting the ICU room. The variation of HAI-relevant events is often attributed to staff fatigue, monotonous routines, or emergencies, and to visitors not being aware of sanitization protocols. The log's objective is to provide a chronological description of events inside the ICU room. The logs can be used by healthcare professionals to backtrack the origins of pathogens, which can help in setting

appropriate corrective plans.

For consistency, assume that events can have one of four qualifiers: clean, contamination, transmission, or unclean. The labels depend on the sequence of underlying actions and the detection of hand sanitation actions performed by people immediately after entering the ICU room and right to the last instance before exiting the ICU room. The variation of HAI-relevant events is often attributed to staff fatigue, monotonous routines, emergencies, or to visitors unawareness of sanitization protocols. The objective of the logs is to provide a chronological description of events inside the ICU room and help understand, minimize, and prevent undesirable sanitation events. The logs can be used by healthcare professionals to backtrack the origins of pathogens, which can help in setting appropriate preventive and corrective action plans.

Identifiable Information and Healthcare Roles. There is an increasing interest in role identification and analysis in healthcare [73], due to its potential benefits in improving and optimizing care. One major gain from role identification and analysis is in defining each person's responsibilities, ensuring appropriate implementation of each professional's role, optimizing professional scopes of practice, and thereby ensuring efficient patient management [7]. Although clinicians agree that detailed understanding of workflows is essential to quality of care, healthcare restrictions prohibit the use of people's identifiable information.

Healthcare Event and Activity Logs. Effective event and activity frameworks such as the one discussed in Chapter 6, require autonomous monitoring systems and algorithms, which are far from being 100% reliable and automated in real-world situations. Such frameworks are required to automatically deal with occlusions, illumination changes, multiple subjects, and concurrent actions and events. Ideal solutions for healthcare

should focus on the detection of human activities and their sequential ordering to classify events into categories (clean or dirty visit). This information can then be used to create sanitation event and visit logs. The major aspects of such solutions are ability to chronologically log events, leverage contextual information, and fuse visual information from multiple views and modalities to deal with natural scene conditions. Solutions must learn and use contextual aspects associated with the various ICU actions. For example, handwashing as an essential activity in healthcare and particularly important for ICU patient health. An approach described in this dissertation identifies and decomposes (and constructs) such an activity into its contextual aspects (roles, duration, objects, etc.).

2.3 Summary of Related Work

This chapter provides some insights on the analysis of patient static and dynamic poses, role representation and identification, and activity and events in healthcare settings. The remaining of this chapter provides a summary of technical related work for each subject previously listed.

2.3.1 Pose Classification

Computer vision methods using RGB data to detect body configurations of patients on beds are discussed in [35, 40, 59] but are limited to scenes with constant illumination and/or without occlusions. The deformable parts model approach, commonly used in RGB images presented in [87] requires images with relatively uniform illumination and is limited to minor self-occlusions. The discriminative approach from [66] uses depth images and is robust to illumination changes. It requires clean depth segmentation and contrast and is susceptible to occlusions. A controlled method to classify human sleep poses using

RGB images and a low-resolution pressure array is presented in [28]. It uses normalized geometric and load distribution features interdependently and requires a clear view of the patient. Also, their system requires complex calibration and a top clear view of the patient's body configuration. Pose classification is also tackled in [77] via RGB, depth, and pressure sensors in simulated healthcare environments. The study in [22] uses bed aligned maps (BAMs) composed of pressure arrays and a single depth camera. Although the BAMs method outperforms previous static sleep pose classification techniques, it does not consider motion. The authors from [75] use convex coupled-constrained least-squares optimization to remove the cumbersome pressure array and create a purely observational system. This latest technique increased the classification accuracy by integrating multi-modal sources from multiple views and creating a truly multiview multimodal sleep pose classification system. Unfortunately, no previous method incorporates time to analyze the sequence of poses, pose transition, or pose motion dynamics. The work in [52] tackles a rehabilitation application via pose detection and tracking; however, its applications are limited to ideal scenarios.

2.3.2 Pose Pattern Analysis

The analysis of human motion dynamics has captured the attention of researchers in the engineering and health communities. In particular, the ailing healthcare system in the U.S. continues to degrade. This degradation requires that engineers and health professionals join forces to develop new efficient therapies and optimize care techniques and workflows. The latest techniques using convolutional neural network (CNN) architectures achieve impressive classification performance. However, CNN-based techniques require large data sets [3], [9], [78], and [79]. In [65], the authors introduced a CNN-alternative method for action representation via sequential deep trajectory descriptors.

The previously cited works recognize actions centered on the camera plane. An exception is the work from [70], which uses stationary cameras and allows off-center actions. A performance limitation of this technique is that it requires scenes with good illumination that are clear of occlusions (i.e., BC). A supervised method for learning local feature descriptors is introduced in [92]. Although effective, the method requires controlled scenes, which are not possible in healthcare. A discriminative multi-instance multitask method to recognize actions in 3D spaces is proposed in [86]. However, this method is unable to distinguish between similar actions, for which their only distinction is their duration.

The work in [69] surveys multimedia methods for large-scale data retrieval and classification using multimedia data. The objective of the survey is to highlight an in-depth understanding of multimedia methods for data analysis and understanding. This will be relevant as the corpus of healthcare monitoring data grows. A true multimedia method to summarize events in videos via audio, visual, and textual saliency is introduced in [13], and a multiview method for surveillance video summarization via sparse optimization are presented in [56]. Although interesting, these methods analyze motion dynamics with less subtlety than the motion of patients in the ICU. Also, these studies analyze scenes with better illumination and are not representative of the ICU environment. In addition, multimedia methods may expect speech or text information as input, which cannot be recorded in the ICU (i.e., hospital). These infrastructural and privacy limitations thwart the implementation and deployment of the existing methods in healthcare applications.

2.3.3 Role Representation and Identification

Studies analyzing healthcare environments include using a single RGB-D sensor, RFIDs, and proximity sensors to record activities in a neo-natal ICU as in [37] Workflows in an operating room are analyzed in [55] and the analysis tasks are very complex. One

significant limitation is considering that any activity can be performed by any individual. This makes the action space relatively large, which decreases accuracy. One helpful concept in improving outcomes includes identifying roles who perform distinctive and common activities and using this information to identify roles (e.g., patient, doctor, or staff). The surveys from [80] and [32] describe the challenges and most popular techniques in person re-identification. Existing methods for identification and re-identification range from methods leveraging deformable parts [8] to feature representation and metric learning [39] to video ranking [82] (as an alternative to single-frame approaches).

The work in [48] introduced a distributed network framework for node performance comparison and person re-identification that can be used to estimate optimal camera topology. The authors in [18] argue that most existing methods depend on person pose and orientation variations and introduce a technique to model such variations in the feature space. Also, there are several feature representations that have pushed the limits of performance to new levels. Appearance-based representations such as the ensemble of local features (ELF) [21] and symmetry-driven accumulation of local features (SDALF) [4] encode color properties. Similarly, salience matching and learning [89], [90] and mid-level filters [91] depend on relative patch contrast and distinctiveness. Although the previously cited research achieves impressive results, their appearance-based methods directly depend on proper imaging conditions, such as bright, and uniform illumination, and view angle between the individual and the camera.

In addition, appearance-based role representation alone is not sufficient. For instance, medical isolation procedures to protect compromised patients require that all people entering the ICU room wear disposable isolation scrubs, so all roles appear identical. Another limitation of these representations in real-world applications, such as healthcare, is their inability to evolve over time (i.e., to consider temporal information) and to integrate interaction information. The proposed approach introduces a novel role representation; a

semantic activity abstraction and extraction algorithm to identify; and a method for role identification based on the sequence of observed activities, visited locations, and detected interactions (cones for orientation and proximity). The proposed methods are capable of dealing with cases when role-based visual features are obfuscated by extreme scene and appearance changes. More details about role identification can be found in [74].

2.3.4 Workflow Analysis: Activity and Event Logging

The latest developments in convolutional neural network (CNN) architectures for action recognition achieve impressive performance; however, these techniques require large data sets [3], [9], [78], and [79]. The methods tackle the recognition of actions performed at the center of the camera plane. The method in [70] uses egocentric cameras to analyze off-center actions. The method in [81] uses CNNs to analyze off-center actions, but it requires scenes with good illumination and clear of occlusions.

Multi-sensor and multi-camera systems and methods have been applied to smart environments [25] and [84]. The systems require alterations to existing infrastructure making their deployment in a hospital logistically impossible. The methods are not designed to account for illumination variations and occlusions and do not account for non-sequential, subtle motion. Therefore, these systems and methods cannot be used to analyze patient motion in a real ICU where patients have limited or constrained mobility and the scenes have random occlusions and unpredictable illumination. In [76] the authors use a multimodal multiview system and combine it with time-series analysis to summarize patient motion. A pose detection and tracking system for rehabilitation is proposed in [52]. The system is developed and tested in ideal scenarios and cannot be used to detect unconstrained motion. In [55] a controlled study focuses on work flow analysis by observing surgeons in a mock-up operating room. The work most similar

to HEAL is introduced in [37], where Radio Frequency Identification Devices (RFIDs) and a single depth camera are used to analyze work flows in a Neo-Natal ICU (NICU) environment. These studies focus on staff actions and disregard patient motion.

An effective solution for activity and event analysis in healthcare must observe the environment and extract contextual aspects from ICU room activities. One such aspect includes identifying people, or when prohibited, identifying roles as in the work from [74]. The events analysis of events requires them to be observed from multiple views and modalities. Their contextual aspects need to be combined with temporal information for proper and accurate representation.

Chapter 3

Static Classification of Patient Sleep Poses

Science is no more than an investigation of a miracle we can never explain, and art is an interpretation of that miracle. -R. Bradbury

3.1 Introduction

Healthcare demands include transforming the healthcare system from disease to patient centric by providing the means to individualize care. For instance, clinical evidence suggests that body poses of patients on beds are correlated with patient recovery rates and response to therapies. The required methods for non-disruptive monitoring and analysis of patient sleep poses, patterns, and quality can add objective metrics for evaluating health-related scenarios.

The standard of care for immobile ICU patients is to rotate them every two hours

to prevent decubitus ulcers, but this is rarely accomplished and has very low compliance rates [68]. The findings in [5, 29, 83] correlate body positions to various effects on health and quality of sleep of ICU patients. The authors state that identification of sleep poses in natural scenarios helps to evaluate sleep and to improve diagnosis and treatment of sleep disorders. Current physiological systems use machines that physically connect to the patients, making them disruptive and intrusive. Purely observational systems use images and pressure arrays to estimate poses but have been unable to handle natural scenes – indoor ICU scenes with variable illumination and occlusions such as blankets and pillows.

Computer vision methods are used in [35, 40, 59] but are limited to ideal scenes. In both approaches, the staging needed for observation affects the measurements. In order to overcome these issues, we propose to use three non-invasive, independent sensor modalities: RGB, depth, and pressure. Existing techniques are able to estimate human poses in ideal scenes using these modalities independently, but they fail in challenging ones. In [87] the authors present a generative approach that uses deformable parts model (DPM), commonly used in RGB images. Unfortunately, the DPM method requires images with relatively uniform illumination and with only minor self-occlusions. The discriminative approach from [66] uses depth images and is robust to illumination changes. However, this method requires clean depth segmentation and contrast, and it fails under occlusions. Neither of these methods works in unconstrained ICU scenarios.

There are two major, connected limitations in the design and dissemination of pose-related healthcare protocols. First, manual analysis is the most effective, but it requires staff to track and record patient poses. Second, automated monitoring systems that can remove the burden from human observers have been unreliable in natural scenarios, which can be partially occluded, poorly illuminated, and continuously changing (e.g., moving equipment). We address these issues by developing methods for robust classifi-

cation of body poses of patients on beds in an Intensive Care Unit (ICU) scenario using multimodal and multiview (mm) data. Our method uses view-adjusted modality trusts – each modality's classification ability under a given set of scene conditions as seen by a particular camera view.

3.2 Systems to Monitor Sleep

Two approaches for sleep pose classification are presented: (1) a multisensor approach, which uses RGB and Depth cameras, environmental sensors, and a complex pressure mat; and (2) a multimodal multiview purely observational approach, which avoids using the complex pressure mat by computing trust values for each sensor modality and view. The two approaches compute trust values and use them to weight pose candidates and infer a final pose for a given observation.

3.2.1 MultiSensor ICU Network

The proposed system shown in Figure 3.3 uses three sensor modalities: a single Carmine device, with standard RGB and depth sensors by Primesense, and a high-resolution, pressure-sensing mattress by Tekscan. The Carmine and Tekscan devices are controlled by DuoCore computers and synchronized using Network Time Protocol (NTP). The computers communicate using TCP-IP. The sensors monitor the bed and actors in a variety of poses and scenes as described in section 3.2.1. The scene context (e.g., illumination level and occlusions) is captured via illumination, proximity, and RFIDs.

The following is a description of the different sensors used to observe the mock-up ICU, with a typical set up is shown in Figure 3.1.

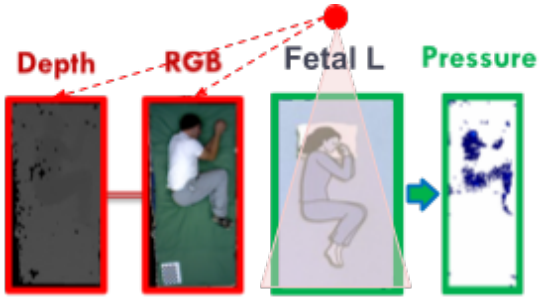


Figure 3.1: Multisensor singleview representation of a sleep pose.
The figure shows the fetal left-oriented sleep pose. The system is composed of one RGB-Depth camera and one pressure mat.

Standard RGB (R). R data provides reliable information to represent and classify patient-on-bed poses in scenes with relatively ideal conditions.

Depth (D). Infrared depth cameras can be resilient to illumination changes. We used Primesense Carmine devices for our depth data collection. These devices are meant for indoor use and can acquire images of dimensions 640×480. The depth-image performance, however, depends on depth contrast, which is affected by the deformation properties of the mattress and blanket. These sensors use 16 bits to represent pixel intensity values, which correspond to the distance from the sensor to a point in the scene. Their operating distance range is 0.8 m to 3.5 m; and their spatial resolution for scenes 2 m away is 3.5 mm for the horizontal (x) and vertical (y) axes, and 30 mm along the depth (z) axis. We used the depth images to represent the 3-dimensional shape of the poses.

Pressure (P). We used the Tekscan Body Pressure Measurement System (BPMS), model BRE5315-4 to collect pressure information and generate pressure images. Although the size of the pressure images is relatively large, the generation and resolution of such images depends on consistent body-mattress contact. In particular, pillows, deformation properties of the mattress, and bed configurations disturb the measurements.

In addition, proper pressure-image generation requires a sensor array with high resolution and full bed coverage, which can be prohibitively expensive and constrictive due to sanitation procedures and requirements.

Why multiple views? Limitations imposed by the pressure mat inspired us to improve multimodal classification performance and to design the proposed MM system and its reduced system configurations. We observe that the classification accuracy is improved when using multimodal information from more than one view of the scene. This is important since real-world scenarios limit modalities and views of the scene (e.g., equipment is moved around the ICU continuously). A multiview system allows us to monitor the bed and patients pose using simple and affordable sensors without requiring expensive pressure mats. The added advantage is that visual sensors (i.e., cameras) are not in contact with the patients, thus avoiding the inherent risk of infections by touch. Finally, though the pressure sensors are generally accurate, they need to be calibrated with respect to the ICU bed configurations, and this is not a trivial process.

Sleep Pose Dataset # 1

Multimodal single bed poses were collected from five actors, who were asked to assume each of the ten poses from set $Z = \{$*Background, Soldier U, Soldier D, Faller R, Faller L, Log R, Log L, Yearner R, Yearner L, Fetal R, Fetal L*$\}$. The set Z has size L and is indexed by l. The letters in the labels U and D stand for facing-Up and facing-Down and L and R stand for laying-on-Left and laying-on-Right. We use z_l to identify a specific pose label (e.g., $z_0 =$ Background). The scene conditions are simulated using three illumination levels: bright (light sensor within 70-90% saturation), medium (50-70%), and dark (below 50%) and four occlusion types: clear (no occlusion), blanket (covering 90% of the actor's body), blanket and pillow, and pillow (between actor's upper body and the pressure

mat). The illumination intensities were assigned using the percent saturation values and the occlusions were detected using inexpensive radio-frequency identification (RFID) and proximity sensors, all by .NET Gadgeteer. The combination of the illumination levels and occlusion types generates a 12-element scene-set $C = \{$(bright, medium, dark) $\times$ (clear, blanket, pillow, blanket+pillow)$\}$. The symbol $c \in C$ is used in the formulation to indicate a single illumination and occlusion combination (e.g., $c = 1$ means bright and clear). The M sensor in $N = \{R, D, P\}$ are calibrated using the methods from [24].

The dataset includes background (bed without actor) images, and images of the actor in each of the 10 poses under each of the 12 scene conditions. The process is repeated ten times for each of the five actors. The dataset contains 26,400 images (5 actors $\times$ 10 sessions $\times$ 4 images $\times$ 11 classes $\times$ 12 scenes). Sample data is shown in Figure 3.2.

Pose Classification Formulation

The sensor trust (w_m^c) is defined as the ability of feature vector f_m for pose classification. The vector f_m is extracted from sensor m under scene conditions c. The trusts are estimated at training, using all the features in the subset X_{train}^c to compare estimated pose label $\hat{z}_k$ to the ground-truth label z_k^* and to record the matches. The learned trusts are used to infer a final multisensor label.

Trust Estimation. The set of sensor trusts $\{w_1, w_2, \ldots w_M\}^c$ is estimated for sensors in N and condition c. The estimation of the trusts is divided into three stages: unisensor training, classifier validation, and trust normalization.

Unisensor Training. The single sensor SVC and LDA classifiers CLF_m^c are trained using the features f_m in X_{train}^c. Each of the classifier outputs a vector $[\hat{s}_{l,k}(f_m)] = [\hat{s}_{1,k}(f_m), \ \ldots, \hat{s}_{L,k}(f_m)]$ of length L. Given a datapoint X_k (with M unisensor feature

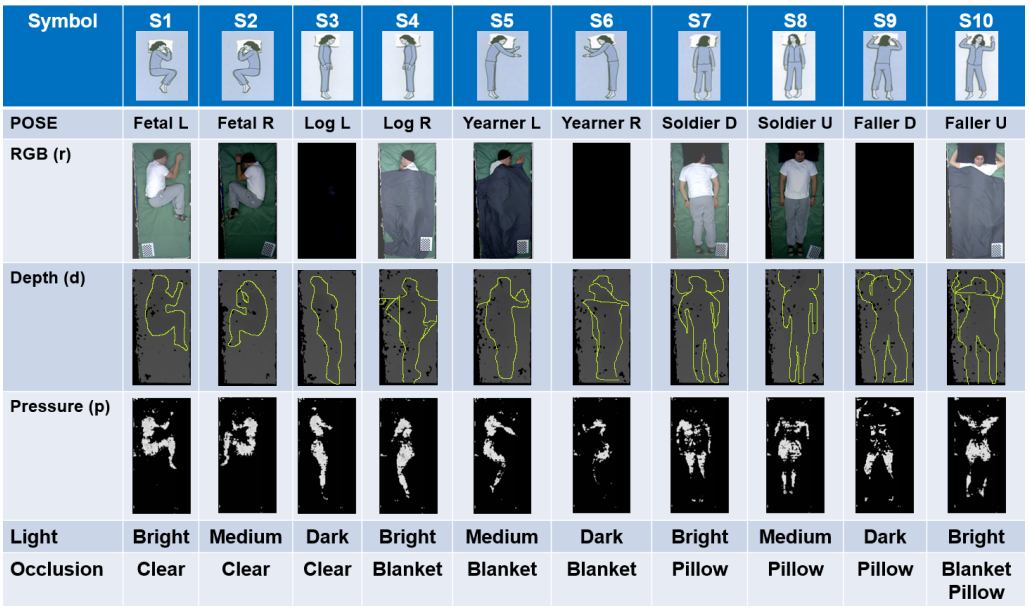


Figure 3.2: Sample multimodal singleview dictionary of sleep poses.
Poses were collected from an actor in various sleep configurations and scenes. The first row is the base pose configuration and orientation, the second row is the pose name, where L and R indicate lying on the Left or Right side, and U and D indicate facing Up or Down. The third, fourth, and fifth rows are the unimodal representations of the pose. The depth images on the fourth row are manually delineated to highlight the differences between the deformable background and the patient's body. Finally, the bottom two rows indicate the scene conditions.

vectors f_m), the $\hat{s}$ elements contain the scores for each of the L labels in Z.

Classifier Validation. At this stage the estimated labels ${Z_{\hat{l},k}}^{(m)}$ are compared to the ground truth label $Z^*_{l,k}$ from data point X_k. The label matches are stored in the array $\mathbf{b}$ of dimensions $[K, M] \forall k$, where ($1 \leq k \leq K$, and $K = |X_{train}|$) using Algorithm 1.

Algorithm 1 Unisensor Classifier Validation Vector ($\mathbf{b}$)

```
1: procedure COMPARE(Z_{l̂,k}^{(m)}, Z*_{l,k})        ▷ Estimated and ground truth labels
2:     b ← 0, m, k = 0,                               ▷ Initialize array
3:     for k do
4:         for m do
5:             if Z_{l̂,k}^{(m)} = Z*_{l,k} then
6:                 b[k, m] ← 1
7:             else
8:                 b[k, m] ← 0
9:             end if
10:        end for
11:    end for
12:    return b                                       ▷ Vector of size K, M
13: end procedure
```

Trust Normalization. Finally, the trusts are estimated with following equation:

$$w_m = \frac{\sum_{k=1}^{K} \mathbf{b}[k, m]}{K}, \tag{3.1}$$

and normalized so that the sum is one.

Pose Classification Formulation. The overall description of the system is shown in Figure 3.5. It uses the single sensor training data (features) to estimate the trust values. Then the system uses the trusts to refine multisensor classification and produce a final pose label for a given test datapoint. First, the system applies a weighted scoring formulation to the unisensor label candidates obtained from the features $\big[f_1, f_2, f_3\big] = \big[\text{HOG}(R),\ \text{gMOM}(D),\ \text{gMOM}(P)\big]$ of datapoint X_k. Finally, the multisensor label $Z_{\hat{l},k}$

is the one with the maximum weighted score as follows:

$$S^c_{m,k} = S^c(X_k[f_m]) = w^c_m \mathrm{CLF}^c_m(f_m), \tag{3.2}$$

where w^c_m represents the predictive power of feature f_m with scene conditions c, and CLF^c_m is the unisensor classifier score vector $[\hat{s}_{1,k}(f_m), \ \ldots, \hat{s}_{L,k}(f_m)]$ (elements are label scores). Thus $S^c_{m,k}$ has L elements representing the label scores for an input X_k. The multisensor score is computed using:

$$S^c_k = \sum_{m=1}^{M} S^c_{m,k} = \sum_{m=1}^{M} \Big(w^c_m [\hat{s}_{1,k}(f_m), \ldots, \hat{s}_{L,k}(f_m)]^c \Big). \tag{3.3}$$

The vector S^c_k has L candidate scores values for each k and is computed via:

$$S^c_k = \sum_{m=1}^{M} \Big(w^c_m \{\hat{s}_{l,k}(f_m)\}^c_L \Big). \tag{3.4}$$

Therefore, given an input vector $X_k = \{f_m\}_M$ from scene c the estimated pose label is $Z_{\hat{l}}$, and the index $\hat{l}$ is computed using the following equations:

$$\hat{l} = \arg\max_{l \in L} \Big(S^c_k \Big), \tag{3.5}$$

where $\hat{l}$ is the index of the label with the highest trusted score from:

$$\hat{l} = \arg\max_{l \in L} \left(\sum_{m=1}^{M} w^c_m \{\hat{s}_{l,k}(f_m)\}^c_L \right). \tag{3.6}$$

3.2.2 Multimodal Multiview ICU Network

This section covers the multimodal multiview sensor network and methods for patient pose classification. In addition, various system configurations are described (pros and cons) and evaluated. The system configurations are based on their active (or missing) modalities and camera views

- Multimodal and Multiview (MM) uses all the resources of the system is composed of the R, D, P modalities and the t, s, h camera views. Experimentally, this configuration has the best performance, is the most complex. It can be expensive and limited to static time analysis due to the use of a highly sensitive, high-resolution pressure mat. Its performance provides an upper bound baseline for the reduced configurations.

- Multimodal partial-Multiview (MpM) uses R, D, P information and less than three camera views $\{t, s, h, ts, th, sh\}$. The individual views were assessed and presented in the experiments section 3.5. This configuration is most similar to the one used in the competing methods, which uses a single top view of the bed scene.

- Partial Multimodal and Multiview (PMM) uses R, D or $R - D$ information from the t, s, and h camera views. Although it does not use information from the pressure mat, its performance is optimal when all views of the scene are usable. This situation is desirable but unrealistic in everyday ICU scenarios where views can be occluded and real-state is limited.

- Partial-Multimodal partial-Multiview ($PMpM$) does not use P information, instead it uses R, D or $R - D$ information from less than three camera views $\{ts, th, sh\}$. With this configuration we seek to ignore the pressure information

using only camera information from two views of the scene. Besides the unimodal single view systems, the $PMpM$ provides the lower bound in performance.

Sleep Pose Dataset # 2

This dataset is collected using a multimodal multiview sensor network and by assuming one scene to be the combination of one actor in one pose and under a single scene condition. From one scene we collected four measurements – three modalities (R, D, and a synthetic binary mask) from three camera views (top: t, side: s, and head: h) and one pressure (P)image. The data collection included background (bed without actor) images, and asking the actor to rotate through the 10 poses (11 classes including the background) under each of the 12 scene conditions. The process is repeated ten times for each of the five actors. The complete process generated a dataset of 66,000 images (5 actors $\times$ 10 sessions $\times$ ((3 views $\times$ 3 images) + 1 pressure image) $\times$ 11 classes $\times$ 12 scenes). Sample data collected using our system is shown in Figure 3.4 for a single actor in various poses and scene conditions. Similarly, sample raw mm data is shown in Figure 3.3 where all views and modalities observe a pose.

Pose Data and Environmental Sensors

The work in MESH is most similar to [28], where standard RGB images and a low-resolution pressure array were used to classify sleep poses from static images. Their method used normalized geometric and load distribution features that depended on a clear view of the scene and the actor. They used interdependent data from RGB and pressure sensors – if one modality failed, no result was produced. Our method uses data from three modalities independently and then combines their estimation results using modality trusts to infer the final pose label. Moreover, our classification method is

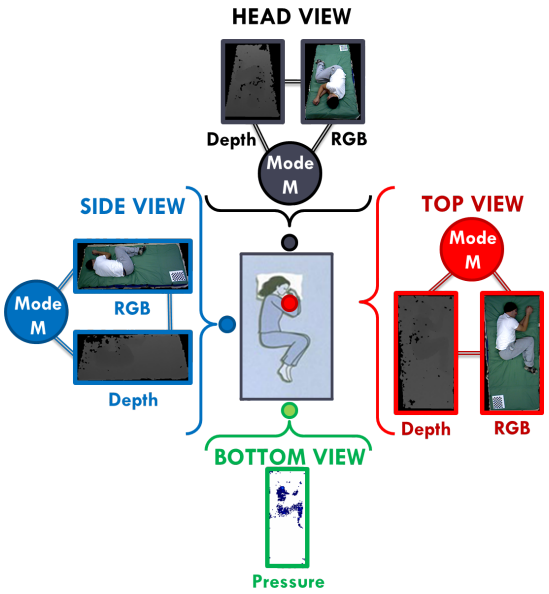


Figure 3.3: Multimodal and multiview representations of a sleep pose.
The figure shows the fetal left-oriented sleep pose. The system is composed of three RGB-Depth cameras and one pressure mat.

independent of body type and we use it to improve the unimodal decision of two common classifiers: Linear Discriminant Analysis (LDA) and Support Vector Classifier (SVC).

Pose Classification via cc-LS

The EYE-CU method uses three modalities – standard RGB (R), Depth (D) and Pressure (P) – in a combinatorial manner to account for variations present in natural scenes. We use three camera views – top (t), side (s), and head (h) – in a complementary way to handle infrastructure occlusions, space limitations, and changes in relative patient-camera orientations. Experimental results indicate that our proposed MM approach matches the performance of existing methods in ideal scenarios – scenes with constant illumination and free of occlusions – and outperforms the latest techniques in challenging clinical scenarios by 13% for those with poor illumination and 70% for those with poor illumination and occlusions. Finally, we demonstrate experimentally that

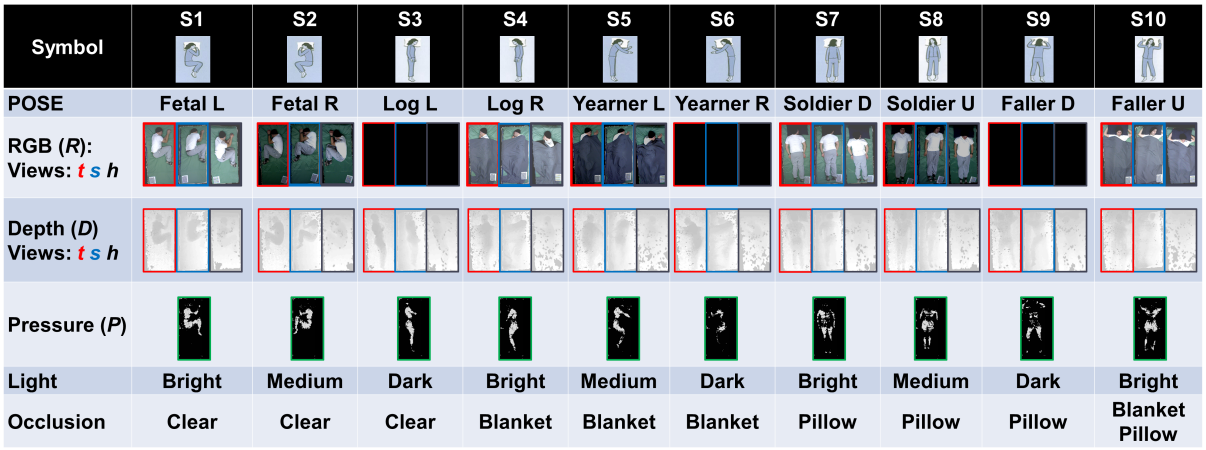


Figure 3.4: Multimodal and multiview dictionary of sleep poses.
Sample poses are from one actor in various sleep pose configurations and scene conditions (illumination levels and occlusion types). The images were generated using R, D modalities from three distinct camera views and one pressure mat P. The images were transformed using the top camera view as the target plane.

variations of our proposed MM method, Partial-Multimodal Multiview (PMM) and Partial-Multimodal partial-Multiview ($PMpM$), are capable of classifying poses without pressure mat information in the same challenging scenarios.

3.3 Feature Extraction

The cameras are first calibrated using standard methods [24] and the corresponding homography transformations are computed. With the top view being the reference plane, these images are then transformed to the reference plane. The following image features are then computed from these transformed images.

Histogram of Oriented Gradients (HOG) HOG feature descriptors are extracted from RGB images. The features use gradients to represent human body pose configurations. The choice is based on [87] where the authors demonstrate the ability of HOG to

represent human limb structures.

Geometric Moments (gMOM) For shape is represented via geometric moments introduced by Hu [27], computed up to the $(i+j)$-th order moment via:

$$\hat{M}_{i,j} = \sum_{x,y} I(x,y) x^j y^i, \tag{3.7}$$

where moment order is given by $i+j$, $i \geq 0$ and $j \geq 0$ are the horizontal (x) and vertical (y) orders, and I is the pixel intensity value (0 or 1 in binary images) at coordinates x, y. The value of $\hat{M}_{0,0}$ represents the area, $\hat{M}_{1,0}$ and $\hat{M}_{0,1}$ represent the centroids of an image. Similarly, the vertical and horizontal moments of inertia are given by $\hat{M}_{2,0}$ and $\hat{M}_{0,2}$. Its abilities for pose shape representation were demonstrated in [1, 61]. In the in-house implementation uses raw pixel values from the tiled depth and pressure images. The tiles use a six-by-six grid from which the moments up to the third geometric moment are extracted (i.e., $\hat{M}_{i,j}$ with $i, j = \{(0,0), (0,1), \ldots, (0,3), (3,0)\}$ for $i+j \leq 3$). Moments computed from each of the 36 blocks generates a 10-element vectors, which are concatenated to generate a 360-element feature vector per image.

3.4 MESH Approach

The multimodal classification framework for a single view system is shown in Figure 3.5. The dataset of poses (X) of size K indexed by k for each of the scene conditions $c \in C$. Explicitly, one datapoint has the form:

$$X_k = \{f_m\}_M = \big[f_1, f_2, f_3\big] = \big[\text{HOG}(R), \text{gMOM}(D), \text{gMOM}(P)\big] \tag{3.8}$$

The HOG and gMOM feature vectors are extracted from their respective modalities in set $N = \{R, D, P\}$ from an observed pose and $M = |N|$. The feature vectors f_m are then used to train unimodal SVM or LDA classifiers (CLF_m). Each of the unimodal classifiers outputs a vector of length L of the form $[\hat{s}_{l,k}(f_m)] = [\hat{s}_{1,k}(f_m), \ \ldots, \hat{s}_{L,k}(f_m)]$. The $\hat{s}$ elements contain the scores for the labels in Z computed for a datapoint X_k with feature vectors f_m.

The feature-classifier combinations are quantified at the trust estimation stage where the unimodal trust values $w^c = \{w_R, w_D, w_P\}$ are computed for condition in c. Intuitively, the term trust can be described as the modality's ability for accurate pose representation and classification under the given scene conditions, which are recorded at training. Finally, the multimodal trusted classifier is constructed by trusting (i.e., weighting) the label decisions of the unimodal ones and combining them into one. The objective of the Multimodal-Multiview formulation is to find the pose label for data point X_k ($\hat{z}_k$) with the highest multimodal-multiview score S. The first step is to compute the single view unimodal scores.

Multimodal Formulation The multimodal formulation allows the system to capture complementary pose data from all the elements in N independently. However, the abilities of the modalities for accurate representation of the pose vary depending on the scene conditions such as dark light or an occluded bed. For example, unimodal trusted scores for modality m for scene with light and occlusion conditions c is represented using $S_{c,m}(z_l)$. For simplicity the scene index c is omitted to get $S_m(z_l)$. The unimodal label scores are computed via:

$$s_m(z_l) = w_m \ CLF_m(f_m), \tag{3.9}$$

where w_m is the trust value, CLF_m is the trained classifier, and f_m is the feature vector

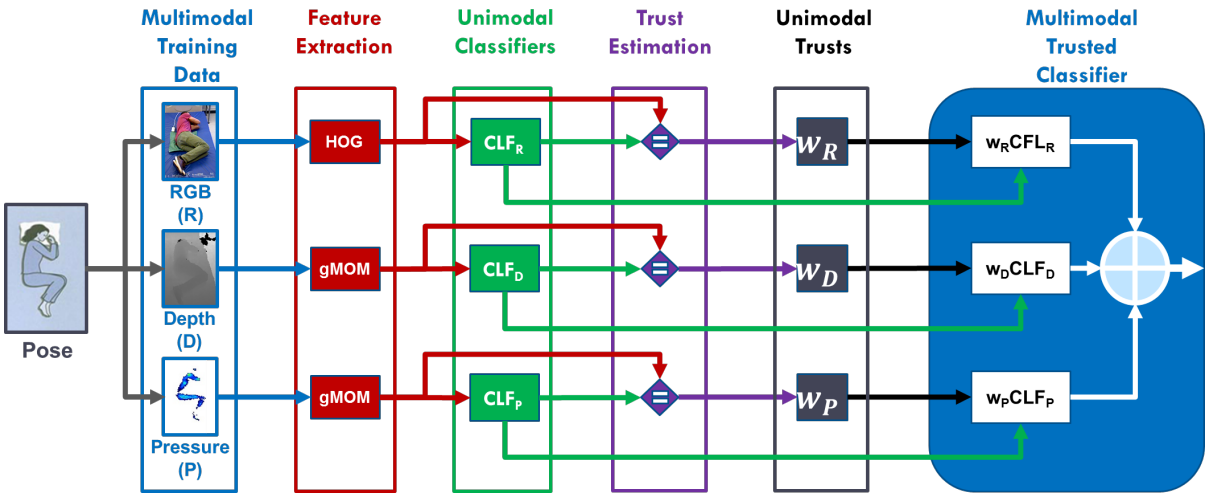


Figure 3.5: Diagram of the trusted multimodal classifier.
The diagram uses the the MpM system configuration data. Features are extracted from the R, D, P pose data using a single camera view and pressure mat. The features are used to train unimodal classifiers (CLF_m), which can be validated as a way to estimate the modality trust values. Finally, in the multimodal trusted classifier stage the unimodal decisions are trusted (i.e., weighted) and combined.

extracted from modality m. The objective is to compute the combined score ($S_k(z_l)$) using data from modalities in N as follows:

$$S_k(z_l) = \sum_{m=1}^{M} s_m(z_l), \forall l. \tag{3.10}$$

An important element for the computation of pose scores is the trust values.

Marginalized Trust Estimation The proposed estimation method for modality trust looks at the ability of each modality for representation and accurate classification of poses using regression and bounded Least-Squares (b-LS). The cc-LS method frames the trust estimation as linear system of equations of the form $\mathbf{Ax}-\mathbf{b}=0$, where the modality trust values are the elements of $\mathbf{x}$. The matrix $\mathbf{A}$ contains the label scores (Z of them) for each of the elements in the training dataset ($K = X_{train}$). The matrix $\mathbf{A} = [\mathbf{A}_1, \ldots, \mathbf{A}_M]$ has LK rows and M columns, where L is the total number of labels ($L = |Z|$), and $M = 3$

is the total number of modalities and has the following structure:

$$\begin{aligned} \mathbf{A} &= \left[s_{k,l,m}\right]_{LK\times 3} \\ &= \left[\mathbf{A}_1, \mathbf{A}_2, \mathbf{A}_3\right]_{LK\times 3} \end{aligned} \tag{3.11}$$

where each column $\mathbf{A}_m$ has the form:

$$\mathbf{A}_m = \begin{bmatrix} \begin{bmatrix} [s_{k,1}(f_m)] \\ \vdots \\ s_{k,L}(f_m) \end{bmatrix}_L \\ \vdots \\ \begin{bmatrix} [s_{K,1}(f_m)] \\ \vdots \\ s_{K,L}(f_m) \end{bmatrix}_L \end{bmatrix}_{KL}, \tag{3.12}$$

where $s_{k,l,m}$ represents the scores for labels Z for datapoint X_k with feature vector f_m.

The column vector $\mathbf{b}$ quantifies the classification ability of the combined modalities. It is used to corroborate estimation of the correct label (when compared to the ground truth training label). The $\mathbf{b}_m$ column vectors are constructed via:

$$\mathbf{b}_m = \begin{bmatrix} b_{k,l} \end{bmatrix}_{LK} = \begin{bmatrix} \begin{bmatrix} \mathbf{b}_{1,l} \end{bmatrix}_L \\ \vdots \\ \begin{bmatrix} \mathbf{b}_{K,l} \end{bmatrix}_L \end{bmatrix}_{LK}, \tag{3.13}$$

with

$$b_{k,l} = \begin{cases} 1, & \text{if } \hat{l} = l^* \\ 0, & \text{otherwise,} \end{cases} \tag{3.14}$$

where $\hat{l} = \arg\max\, \hat{s}_{l,k}(f_m)$ is the index of the estimated pose label and l^* is the index of

the ground truth label for data point X_k.

Finally, the construction of $\mathbf{b}$ depends on how the $\mathbf{b_m}$ column vectors are combined. The three basic construction methods are tested: uniform, at-least-one, and modality consensus. Findings are reported using uniform (the best results) in section 3.5. In the uniform scheme, each modality has a $1/M$ voting power. Each modality proposes a candidate label and modalities are to accumulate votes and its construction is given by the following averaging equation:

$$\mathbf{b} = \frac{\sum_{m=1}^{M} \mathbf{b_m}}{M}. \tag{3.15}$$

In the at-least-one scheme, each modality can propose different label candidates. It functions as an OR logical operator in which a candidate label has a value of one when selected (or activated) and a value of zero otherwise. Votes do not accumulate but serve to activate pose label candidates as follows:

$$\mathbf{b} = \bigcup_{m=1}^{M} \mathbf{b_m} \tag{3.16}$$

The modality-consensus scheme is a greedy approach that provides a single candidate label. It functions as the AND logical operator: each modality assigns, at most, one vote an only the candidate with M votes is selected and its construction is given by:

$$\mathbf{b} = \prod_{m=1}^{M} \mathbf{b_m}. \tag{3.17}$$

Experimentally, allowing $\mathbf{b}$ to accumulate votes (i.e., equation (3.15)) provides the best classification results overall.

Finally, the weight vectors $\mathbf{x} = [w_R, w_D, w_P]^T$ are estimated by solving the constrained

optimization problem:

$$\underset{x}{\text{minimize}} \frac{1}{2}\|\mathbf{A}\mathbf{x} - \mathbf{b}\|^2 \tag{3.18}$$

subject to $\mathbf{0} \leq \mathbf{x} \leq \mathbf{1},\ \mathbf{1}^T\mathbf{x} = 1$.

Intuitively, the solution given found via Least-Squares problem provides priors for every modality.

Multiview Formulation Each modality and each view is capable of acquiring body-pose information. We expand the bounded multimodal formulation to incorporate multiview information for a system w V views indexed by v. The $\mathbf{A}$ matrix is expanded to include the multivew information such that $\mathbf{A} = [\mathbf{A}^{(1)}, \ldots, \mathbf{A}^{(V)}]^T$. Each view-dependent $\mathbf{A}$ is constructed as follows:

$$\mathbf{A}^{(v)} = \left[\hat{s}_{l,k}^{(v)}(f_m)\right]_{LKV \times M} \tag{3.19}$$

where v is the view index in a system with V views.

Similarly, the column vector $\mathbf{b}_m$ is constructed by concatenating information from each view ($v \in V$) such as:

$$\mathbf{b}_m = \begin{bmatrix} [b_{KL}]_{v=1} \\ \vdots \\ [b_{KL}]_{v=V} \end{bmatrix}_{KLV} \tag{3.20}$$

Finally, the $\mathbf{b}$ column vector is computed using equation (3.15).

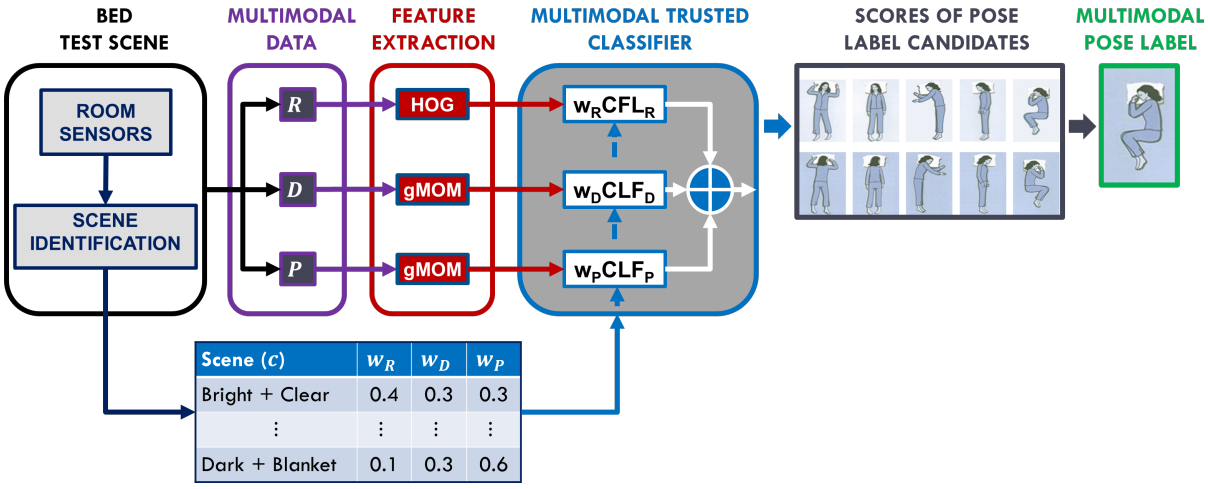


Figure 3.6: Evaluation diagram for the multimodal singleview trusted classifier. Observations are taken from a bed scene (R, D, P) from which features are extracted. Then the features are sent to the multimodal trusted classifier block. Inside this block each of the trained classifiers provides a set of score-ranked pose candidate labels. Finally, the set of unimodal candidate label scores are weighted and combined into one set. From this multimodal candidate set, we select the candidate with the highest score.

3.4.1 Testing

The test framework is shown in Figure 3.6. The room sensors (illumination and RFIDs) in combination with $N = \{R, D, P\}$ data measurements are collected from the ICU room scene. Features (f_m) are extracted from the N modalities and used as inputs to the multimodal classifier.

The final step requires calculating and ranking the pose candidate labels and selecting the label with the largest probability. For example, given a condition c and a data point k, the estimated label is represented by $\hat{z}_k$ and is the candidate label with the maximum score $S_c(z)$ and given by:

$$\hat{z}_k = \arg\max_{z \in Z} \left(S_c(z) \right). \tag{3.21}$$

3.5 MESH Experimental Results

Reported results are based on two classification methods: multi-class linear SVC and LDA from [57] with a five-fold cross-validation scheme for all reported accuracies. Experiments show that illumination affects R performance, while the performances of D and P remain constant. Recognition using R and D is affected by visual occlusions (blankets) and P is affected by pillows.

Unisensor. The system is trained and tested with a single concatenated vector (RDP), and the unimodal vectors R, D, P. The experiments are shown in Table 3.1. These results provide a performance basis for classification and justify the need for a multimodal approach. The results suggest that neither the concatenation of all nor the use individual modalities recognizes poses in all scenes.

Multisensor. These experiments show that the system considerably classifies sleep poses in ICU scenarios using modality trust. Extensive literature search indicates that there is no other method that considers the range of scenarios explored these experiments. The performance of the proposed method is shown in Table 3.2 and compared to Majority-Vote-Learner (MaVL) for contrast and with an in-house implementation of [28]

Missing Sensors. The limits of the system is tested by omitting one sensor at a time and adjusting the contributions of the remaining sensors. Results are reported for SVC and LDA accuracies for all considered scenes in Table 3.3. Results indicate that the system performs poorly without pressure information, achieving a classification accuracy of 6% using SVC and 12.8% using LDA for dark and occluded scenes. The adjusted value of a missing sensor (n) is set to zero (i.e., $w_n^* = 0$), and the adjusted trust values of the remaining sensors (w^*) are computed by proportionally distributing the value of

SCENE		PROPOSED-SVC				PROPOSED-LDA			
Illumination	Occlusion	*RDP*	*R*	*D*	*P*	*RDP*	*R*	*D*	*P*
Bright	All Occlusions	81	79	20	64	52	81	82	94
	Clear	75	100	23	55	99	100	97	93
	Blanket	60	70	22	55	83	79	78	93
	Blanket+Pillow	58	66	24	56	80	76	73	80
	Pillow	66	76	25	56	85	80	75	80
Medium	All Occlusions	72	79	20	64	51	52	82	94
	Clear	88	100	25	55	100	100	98	94
	Blanket	67	70	22	55	83	79	76	94
	Blanket+Pillow	78	66	24	56	90	76	72	83
	Pillow	62	74	25	56	85	83	73	80
Dark	All Occlusions	62	10	21	64	85	9	80	86
	Clear	67	22	25	55	75	12	97	94
	Blanket	55	9	22	55	70	5	78	94
	Blanket+Pillow	57	9	24	56	72	5	75	80
	Pillow	56	9	25	56	80	5	75	84
All Lights	All Occlusions	51	64	21	74	70	53	65	70
	Clear	82	72	82	83	49	52	88	100
	Blanket	43	50	22	63	67	37	55	76
	Blanket+Pillow	31	37	14	60	66	40	51	72
	Pillow	53	60	25	60	77	54	72	69

Table 3.1: Unisensor sleep-pose classification accuracy.
The unimodal mean classification accuracy uses the feature vectors (R, D, P) and concatenated modalities (RDP) used without modality trust for SVC and LDA classifiers for all illumination levels and individual occlusion types. Results suggest that a system that directly uses these modalities for classification is unreliable.

the missing (w_n) using the following equations:

$$w_m^* = w_m \left(1 + \frac{|w_n - w_m|}{W}\right) \text{ for } m \in \{1, 2, \ \ldots, M\} \setminus n, \text{ where} \tag{3.22}$$

$$W = \sum_{m=1}^{M} w_m. \tag{3.23}$$

Multisensor Confusion Matrices. Matrices shown in Figure 3.7 are constructed using the multisensor and [28] methods. The main diagonal on the right shows that the multisensor method outperforms the competition in natural scenes.

SCENE		COMPETING		PROPOSED	
Illumination	Occlusion	MaVL (RDP)	Huang (RP)	**SVC** (RDP)	**LDA** (RDP)
Bright	All Occlusions	75	73	99.3	98.7
	Clear	80	100	100	100
	Blanket	82	8	85.8	80.4
	Blanket+Pillow	65	6	85.8	83.6
	Pillow	54	58	90	90
Medium	All Occlusions	76	70	89.3	88.7
	Clear	80	88	100	100
	Blanket	65	7	85.3	80.6
	Blanket+Pillow	57	7	85.3	83.6
	Pillow	78	37	90	90
Dark	All Occlusions	45	–	24.6	33.3
	Clear	17	–	81.2	85
	Blanket	20	–	20.0	19.2
	Blanket+Pillow	32	–	17.7	18.6
	Pillow	60	–	24.5	22.3
All Lights	All Occlusions	50	–	71.5	85.2
	Clear	75	–	82.6	94.1
	Blanket	54	–	70.2	69.8
	Blanket+Pillow	48	–	67.3	66.5
	Pillow	68	–	77.3	79.8

Table 3.2: Multisensor sleep-pose classification accuracy.
Mean classification methods of two competing methods and our proposed trust using SVM (SVC) and LDA classifiers. Multisensors match the performance of two competing methods in bright, clear scenes and outperforms them approximately by 30 to 50% in challenging scenes (see Dark block).

3.5.1 Performance of cc-Ls Classification

Validation of modalities and views for sleep-pose classification substantiates the need for a multiview and multimodal system. The cc-LS method is tested on the MpM, MM, PMM and $PMpM$ Eye-CU configurations and data collected from scenes with various illumination levels and occlusion types. The labels are estimated using multi-class linear SVC (C=0.5) and LDA classifiers from [57]. A validation set is used to tune the SVC's C parameter and the Ada parameters. Classification accuracies are computed using five-fold cross validation using in-house implementations of competing methods and reported as percent accuracy values inside color-scaled cells.

Classification results obtained using unimodal and multimodal data without modality trust are shown in Figure 3.8. The cell values indicate classification percent accuracy for

SCENE		SVC			LDA		
Illumination	Occlusion	$RD \setminus P$	$RP \setminus D$	$DP \setminus R$	$RD \setminus P$	$RP \setminus D$	$DP \setminus R$
Bright	All Occlusions	89.3	89.3	100	88.4	88.7	88.7
	Clear	100	100	100	100	100	100
	Blanket	85	90	95	80	85	92
	Blanket+Pillow	80	85	90	88.6	83.6	83.6
	Pillow	85	88	87	90	85	95
Medium	All Occlusions	79.3	89.3	89.3	88.5	88.7	88.7
	Clear	100	100	100	100	100	100
	Blanket	70	80	75	68.6	78.6	88.6
	Blanket+Pillow	65	70	71	73.5	81.6	83.6
	Pillow	81	85	87	77.3	82	85
Dark	All Occlusions	25.8	11.3	24.2	37.8	70.9	81.3
	Clear	54.1	47.7	72.7	29.5	74.1	76.4
	Blanket	7.5	30	35	23.2	68.6	76.8
	Blanket+Pillow	6	27	30	12.8	53	68.9
	Pillow	12	37	45	36.3	65.1	73.7
All Lights	All Occlusions	62	65	72	58.1	61.8	75.2
	Clear	62.3	61.8	77	55	22.4	82.7
	Blanket	50	57	72	50	29.2	23.8
	Blanket+Pillow	47.3	54.3	67.3	49.8	52.1	66.5
	Pillow	61	67	78	59.5	70	70

Table 3.3: Classification accuracy with incomplete multisensor information.
The results are mean classification accuracy with one modality omitted (\) and the trust values of the remaining modalities are adjusted for SVC and LDA.

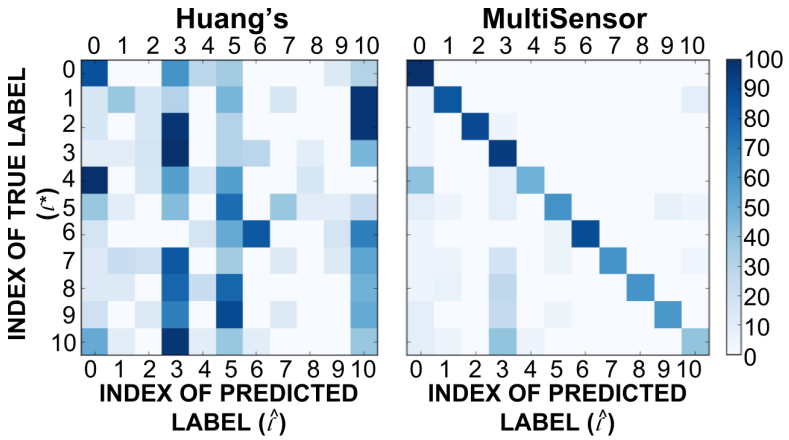


Figure 3.7: Multisensor and Competing confusion matrices.
The implementation of [28] achieves a 16% and our proposed method achieves 70% for dark and occluded scenarios. The confusion matrices show how the indexes of the estimated labels $\hat{l}$ match the actual labels l^*. Main diagonal indicates that the multisensors method (right) performs better.

each individual modality and modality combinations with three common classification methods. The labels of the column blocks at the top of the figure indicate modalities used, while the bottom labels indicate the classifier used. The labels on the left and right indicate scene illumination level and type of occlusion. The figure shows classification results for the top camera view because variation across views tested did not have statistical significance.

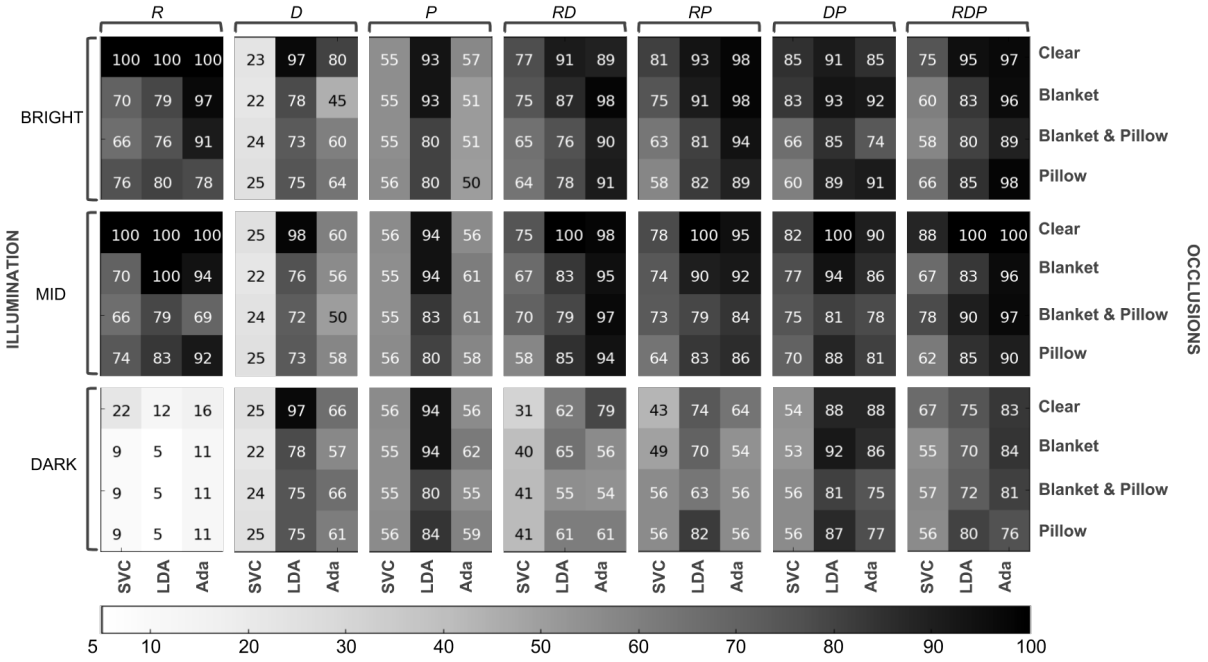


Figure 3.8: MESH performance evaluation of modality combinations.
This evaluation uses SVC, LDA, and Ada-Boosted SVC (Ada) based on their classification percent accuracy (cell values). The evaluation is performed over all the scene conditions considered in this study. The results indicate that no single modality (R, D, P) or combination of concatenated modalities (RD, RP, DP, RDP) in combination with one of three classification techniques cannot be directly used to recognize poses in all scenes. The top row indicates which modality or combination of modalities is used. The labels on the bottom indicate which classifier is used. The labels to the left and right indicate the scene's illumination level and occlusion types. The gray-scaled boxes range from worst (white) to best (black) performance.

3.5.2 Reduced System Configurations

The complete MM configuration achieves the best classification performance, followed closely by the performances of the MpM, PMM, and $PMpM$ configurations, which is summarized in figure 3.9. The values inside the cells represent classification percent accuracy of the cc-Ls method combined with various Eye-CU system configurations. The top row indicates the configuration. The second row indicates the views. The labels on the bottom of the figure identify the modalities. The labels on the left and right indicate illumination level and occlusion type. The red scale ranges from light red (worst) to dark red (best). The figure shows that the complete MM system in combination with the cc-LS method performs the best across all scenes. However, it requires information from a pressure mat. The PMM and $PMpM$ configurations do not require the pressure mat and are still capable of performing reliably and with only a slight drop in their performance. For example, in dark and occluded scenes the PMM and $PMpM$ configurations reach 77% and 80% classification rates, respectively (lower block: DARK; row: Blanket & Pillow) shown in Figure 3.8.

3.5.3 Comparison with Existing Methods

Performance of the cc-LS and the in-house implementations of the competing methods from [28] and [77] and Ada [17] are shown in Figure 3.10. The figure shows results using the MpM configuration, which more closely resembles those used in the competing methods. All the methods use a multimodal system with a top camera view and a pressure mat. The values inside the cells are the classification percent accuracy. The green scale goes from light green (worst) to dark green (best). The top row divides the methods into competing and proposed. The second row cites the methods. The bottom row indicates which classifier and, in parentheses, modalities are used. The labels on the

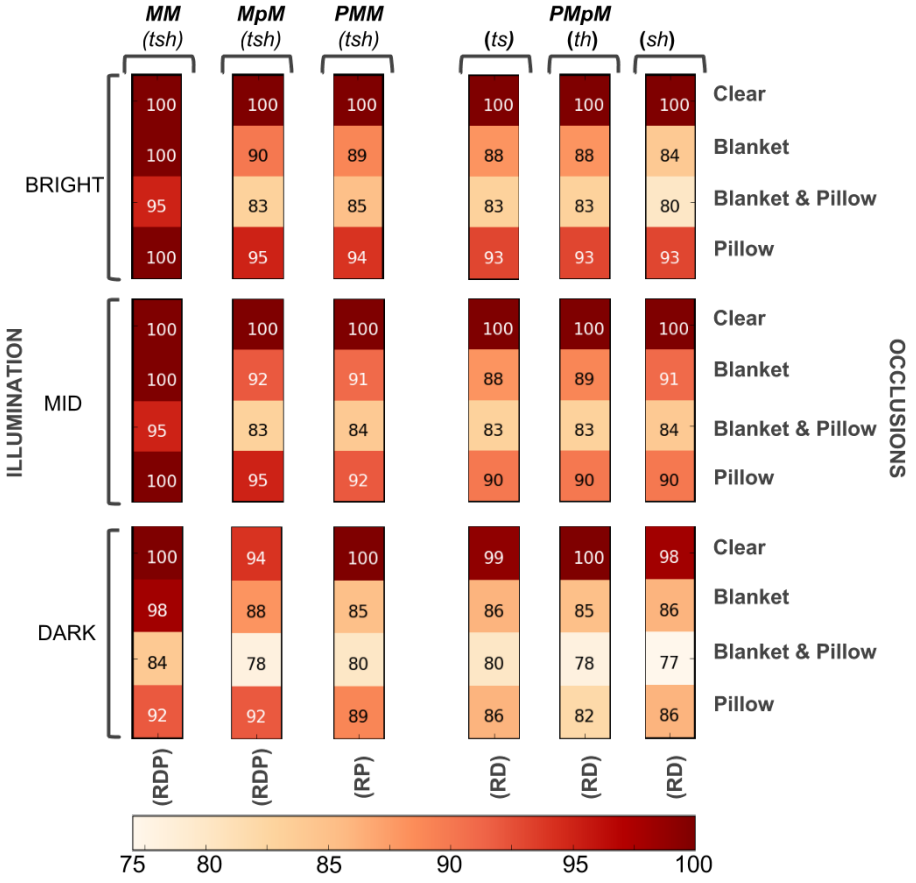


Figure 3.9: Pose classification performance based on system configuration. Performance in red scale (dark: best, light: worst) of the various Eye-CU configurations using LDA. The $PMpM$ has the lowest performance of 76.7% using sh views of a dark and occluded scene. The method from [77] performs below 50% and the method from [28] is not suited for such conditions. The top row identifies the configuration. The second row indicates views used. The bottom labels indicate modalities used (in parenthesis). The labels on the left and right indicate scene illumination and occlusion type. Similar pattern is observed in SVC.

left and right indicate illumination level and occlusion type. The results are obtained using the four methods with MM dataset.

Confusion Matrices: The confusion matrices in Figures 3.11 and 3.12 show how the indexes of estimated labels $\hat{l}$ match the actual labels l^*. The top three matrices are from a scene with bright and clear ICU conditions (Figure 3.11). The bottom three matrices

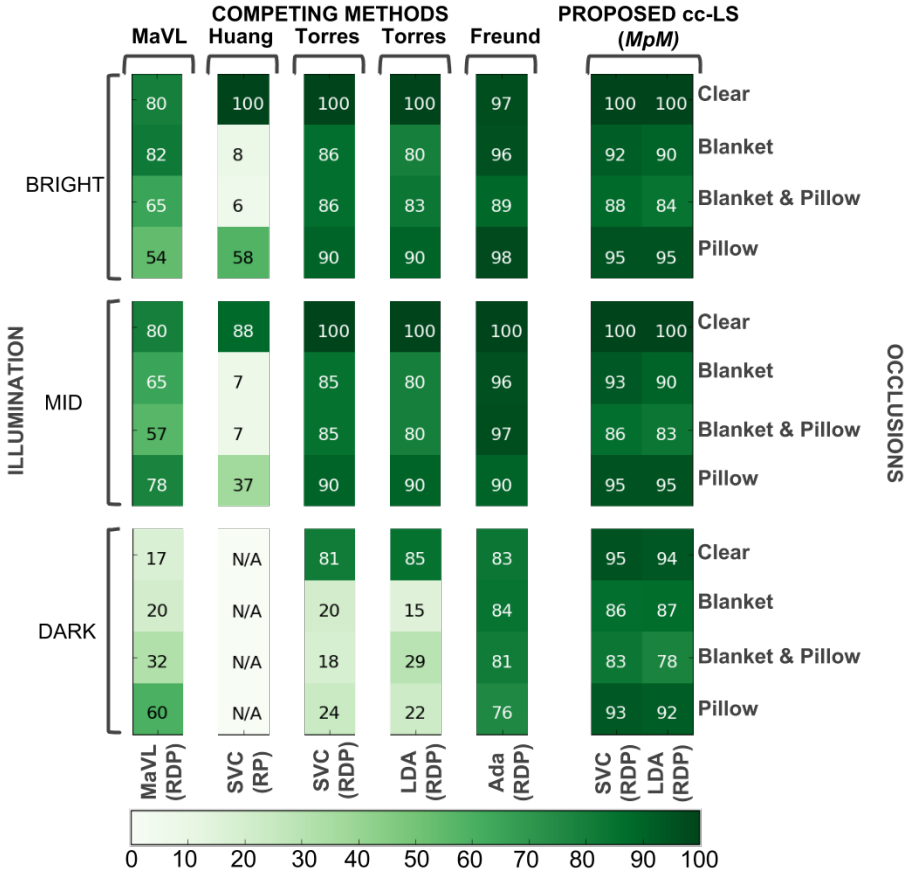


Figure 3.10: MESH mean classification performance comparison.
Performance in green scale (dark: best, light:worst) of MaVL, Huang's [28], Torres' [77], Feund's [17] and the cc-LS method using SVC and LDA. The combination of cc-LS and MpM matches the performance of competing methods in bright and clear scenes. Classification is improved with cc-LS by 70% with SVC and by 30% with LDA in dark and occluded scenes. The top row distinguishes between competing and proposed methods; the second row cites them. The bottom row indicates classifier and modalities (in parenthesis) used. The labels on the left and right indicate scene illumination and occlusion type. N/A indicates not suitable.

illustrate the performance of the methods in a dim and occluded ICU scenario (Figure 3.12). A dark blue diagonal in the confusion matrices indicates perfect classification. In the selected scenes, all methods achieved a 100 % classification for the bright and clear scene. However, their performance varies greatly in dim and occluded scenes. The matrix generated using [28] achieves 7% classification accuracy (bottom left), matrix generated

using [77] achieves a 55% accuracy (bottom center), and the matrix generated with the cc-LS method achieves a 86.7% accuracy (bottom right). The MpM configuration with the cc-LS method outperforms the competing methods by an approximate 30%.

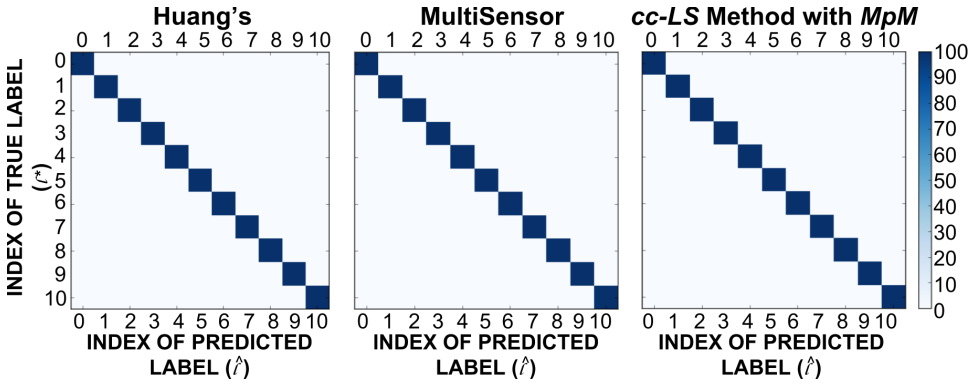


Figure 3.11: MESH pose confusion matrices of three methods in BC scenes. The matrices are generated in blue scale (dark: best, light: worst) using a top camera view and applying the methods from Huang's [28], Torres' [77], and cc-LS with MpM. The matrices show all methods have perfect classification in BC scenes (i.e., main diagonal).

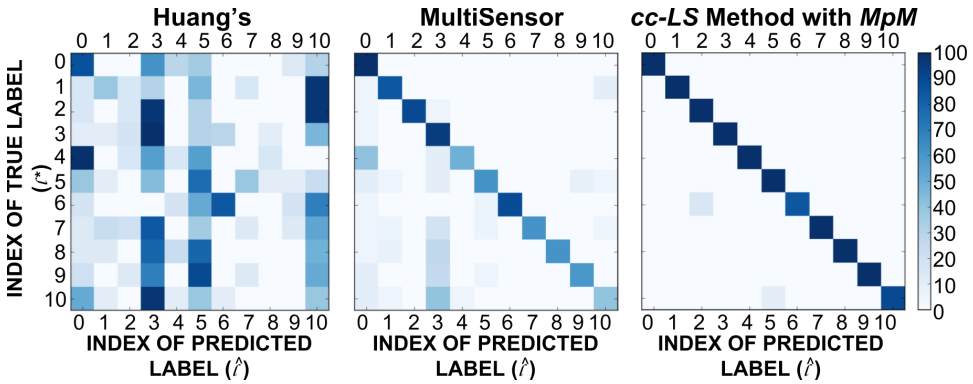


Figure 3.12: MESH pose confusion matrices of three methods in DO scenes. The matrices are generated in blue scale (dark: best, light: worst) using a top camera view and applying the methods from Huang's [28]with 7%, [77] with 55%, and cc-LS with 86.7% for dark and occluded scenes. The matrices show the matches between estimated ($\hat{l}$) and ground truth (l^*) indices.

Performance of Ada-Boost The system is tested using Ada-Boost (Ada) algorithm [17] to improve the decision of weak unimodal SVCs. The results from Figure 3.8 show a slight SVC improvement. The comparison in Figure 3.10 shows that the Ada's improvement is small. It barely outperforms the reduced MpM configuration with cc-LS method in some scenes (see row: MID-Blanket). Overall, Ada is outperformed by the combination of cc-LS and MpM.

3.6 Summary

This Chapter introduced a multisensor singleview and a multimodal multiview sensor network for healthcare along with methods and algorithms to reliably classify sleep poses in ICU settings. The multisensor network uses a complex pressure array and single view RGB-Depth camera sensor. The multimodal mutiview network can be configured to be purely observational and to avoid using the pressure array by incorporating additional views. The various sensors and features used in both generations are thoroughly evaluated for sleep pose classification in healthcare. Also, this Chapter described methods to construct a multimodal multiview matrix and an oracle vector. These elements are strategically combined to solve a cc-LS optimization problem and estimate modality and view trust values.

Quantitative experimental results show that the system has a performance increase of 6% with respect to the two existing methods in ideal scenarios and outperforms them significantly in dark and occluded ones. Reliability of the methods is tested by sequentially omitting information from one modality and adjusting the remaining ones, which achieves a classification accuracy of 47% in challenging scenes. In addition, being able to classifying poses of on-bed patients without the use of a pressure mat (hence avoiding sanitation and integrity concerns), the proposed cc-LS solution accounts for scene and

sensor variations.

The multimodal multiview cc-LS method reliably classifies sleep poses in natural static ICU scenes; however, it does not use temporal information, which is essential to sleep-pattern analysis. The next Chapter 4 introduces new methods to incorporate temporal information and analyze pose patterns: pose sequences, pose transitions, and transition quantification (direction and range).

Chapter 4

Dynamic Analysis of Patient Pose Patterns

We have doomed the wolf not for what it is, but for what we deliberately and mistakenly perceive it to be. -F. Mowat

4.1 Introduction

While receiving care in hospitals, ICU patients are continuously monitored by health-care staff; however, there are no clinical procedures to reliably analyze and understand pose variations from observations (e.g., videos) or the effects of time-based pose patterns on patient health. The recovery of ICU patients varies largely and often inexplicably [19], even for patients with similar initial health conditions. A small number of clinical studies [50] suggests that patient therapies based on body positioning and controlled motion can enhance patient recovery, while inadequate positioning can have negative effects

and aggravate patient health. This study attempts to address this crucial healthcare deficiency by introducing MASH's algorithms and multimodal multiview (mm) camera network. MASH is an autonomous system, which addresses these issues by monitoring healthcare environments and enabling the recording and analysis of patient sleep-pose patterns. MASH uses three RGB-D cameras to monitor patients in an ICU room. The proposed algorithms estimate pose direction at different temporal resolutions and use keyframes to efficiently represent pose transition dynamics. MASH combines deep features computed from the data with a modified version of Hidden Markov Model (HMM) to flexibly model patient pose duration and summarize patient motion.

The MASH architecture analyzes input videos from multiviews and modalities to deal with variable scene conditions from a purely observation approach. Motion quantization is performed to remove depth's sensor noise and threshold observable levels of detectable motion. After noise and motion thresholding, features are extracted to represent the various poses and pseudo or transitory poses (deep and/or engineered features). MASH uses keyframes because collecting, storing, and processing video data from the six sources becomes a hefty task on its own. This problem is more manageable using keyframes across all views and modalities, which can be considered as the frames that are informative and discriminant (i.e., pose and pseudo-pose centroids). Pose patterns and pose transitions can span seconds, minutes, or hours, so we use a modified HMM to flexibly model state or pose duration. Finally, the summary can tell us whether the observation was a sequence of poses seen over an extended period of time (i.e., hours) or the same sequence of poses a transition (i.e., seconds). With these considerations, the workflow shown in Figure 4.1 consists of six major blocks: (1) data collection regarding sleep poses and pose transitions; (2) motion thresholding, which uses optic flow vectors to remove noise from the depth cameras and subtlety distinguish between small and large movements; (3) features extracted from the last layer of the Inception architecture [71] to represent

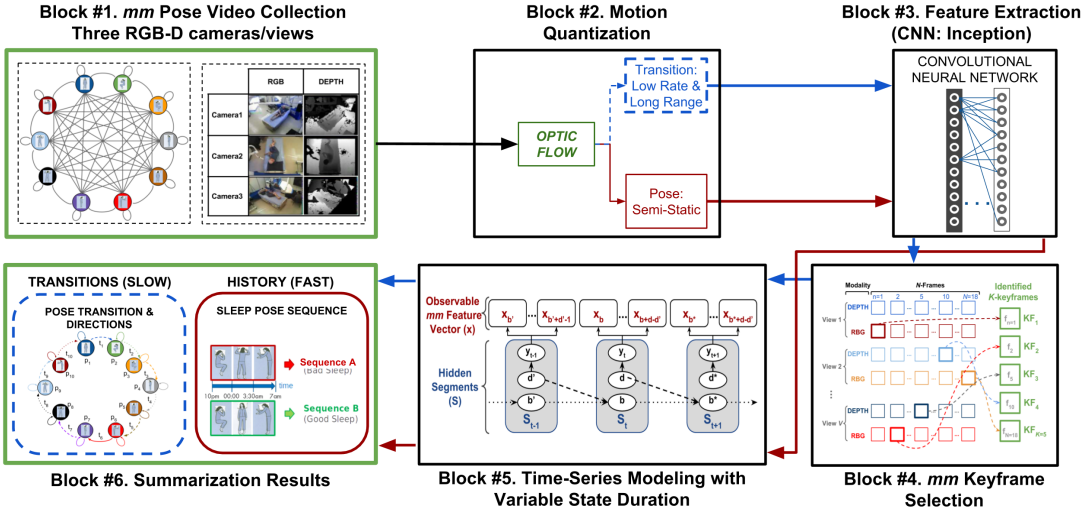


Figure 4.1: MASH framework blocks.
The process starts with Block #1 and flows clock wise: data collection, motion threshold, deep feature extraction, *mm* keyframe selection, time-series modeling, and inferred summarized results.

body configurations as a numerical vector, (4) keyframe extraction to identify pseudo poses that best represent a transition; (5) time-series analysis via HSMM to identify the most likely sequence and model pose duration; and (6) output summary.

The performance is evaluated in ideal (BC: Bright and Clear/occlusion-free) and natural (DO: Dark and Occluded) scenarios at two motion resolutions and in two environments: a mock-up and a medical ICU. The usage of deep features is evaluated and their performance compared with engineered features. Experimental results using deep features in DO scenes increases performance from 86.7% to 93.6%, while matching the classification performance of engineered features in BC scenes. The performance of MASH is compared with HMM and C3D. The overall over-time tracing and summarization error rate across all methods increased when transitioning from the mock-up to the the medical ICU data. The proposed keyframe estimation helps achieve a 78% transition classification accuracy.

Technical Background. The analysis of human motion dynamics has captured the attention of researchers in the engineering and health communities. In particular, the ailing healthcare system in the U.S. continues to degrade. This degradation requires that engineers and health professionals join forces to develop new efficient therapies and optimize care techniques and workflows. The latest techniques using convolutional neural network (CNN) architectures achieve impressive classification performance. However, CNN-based techniques require large data sets [3], [9], [78], and [79]. In [65], the authors introduced a CNN-alternative method for action representation via sequential deep trajectory descriptors. The previously cited works recognize actions centered on the camera plane. An exception is the work from [70], which uses stationary cameras and allows off-center actions and is limited to scenes with good illumination that are clear of occlusions (i.e., BC). A supervised method for learning local feature descriptors is introduced in [92]. Human action recognition benchamarks are described in [42] with best practices for recognition outlined in [58] and [85]. The spatio-temporal evolution of features for action recognition is explored in [43] and [41]. However, activity and action motion observed in conventional activity recognition problems with events such as walking and running. Sleep-pose patterns are different; they are subtle, non-continuous, non-sequential, and abrupt. Although effective, the method requires controlled scenes, which are not possible in healthcare. A discriminative multi-instance multitask method to recognize actions in 3D spaces is proposed in [86]. However, the proposed method is unable to distinguish between similar actions, for which their only distinction is their duration.The ICU scenes and bed setting disqualify techniques based on skeletal estimation and tracking [2] and pure RGB data for human body orientation [44]. Although promising, the work described in [45] is limited by partial occlusions and challenging ICU bed configurations, which are tackled using multimodal multiview data.

The work in [69] surveys multimedia methods for large-scale data retrieval and clas-

sification using multimedia data. The objective of the survey is to highlight an in-depth understanding of multimedia methods for data analysis and understanding. This will be relevant as more data is collected by MASH. A true multimedia method to summarize events in videos via audio, visual, and textual saliency is introduced in [13], and a multi-view method for surveillance video summarization via sparse optimization are presented in [56]. Although interesting, these methods analyze motion dynamics with less subtlety than the motion of patients in the ICU. Also, these studies analyze scenes with better illumination and are not representative of the ICU environment. In addition, multimedia methods may expect speech or text information as input, which cannot be recorded in the ICU (or hospital space). These infrastructural and privacy limitations thwart the implementation and deployment of the existing methods in healthcare applications. The studies from [25] and [84] use multiview systems and methods for smart environments. Unfortunately, these methods require modifications to existing infrastructure. The aforementioned studies are limited to ideal scenes because they cannot overcome illumination variations and occlusions. Furthermore, they do not account for subtle motion, which can be non-uniform and non-sequential. Hence, these cannot be deployed in a medical ICU where systems and techniques are required to be autonomous, non-intrusive, and non-disruptive. These cannot be used to analyze motion in the ICU where patients have limited mobility.

The authors from [28] introduced an RGB-pressure system for sleep pose classification. Their technique uses geometric features to represent poses extracted from the pressure array and the static RGB image. However, the system requires complex calibration and a top clear view of the patient's body configuration. Pose classification is also tackled in [77] using RGB, depth, and pressure sensors in simulated healthcare environments. The authors combine RGB, depth, and pressure modalities with room sensors to weight modality reliability. The study in [22] uses bed aligned maps (BAMs) composed of

pressure arrays and a single depth camera. Although the BAMs method outperforms previous static sleep pose classification techniques, it does not consider motion. The authors from [75] use convex coupled-constrained least-squares optimization to remove the cumbersome pressure array and create a purely observational system. This latest technique increased the classification accuracy by integrating multimodal sources from multiple views and creating a truly multiview multimodal sleep pose classification system. Unfortunately, no previous method incorporates time to analyze the sequence of poses, pose transition, or pose motion dynamics. The work in [52] tackles a rehabilitation application via pose detection and tracking; however, its applications are limited to ideal scenarios. MASH observes the environment from multiple modalities and multiple views to account for challenging natural scene conditions. Two distinctive aspects include: incorporation of variable time information and ability to deal with subtle motion patterns using principled statistics.

Contributions. The technical contributions of this work are: (1) an adaptive framework capable of monitoring patient motion at various resolutions; (2) a non-disruptive and non-obtrusive monitoring system robust to natural healthcare scenarios and conditions such as variable illumination and partial occlusions; (3) an algorithm that effectively compresses sleep pose transitions using a subset of the most informative and most discriminative keyframes; and (4) a fusion technique to incorporate observations from multiple modalities and views (complementary data) into emission probabilities to estimate intermediate poses and pose transitions over time.

4.2 System

MASH is a multimodal multiview (*mm*) framework to monitor patients in healthcare environments independent of motion rate and range. Its elements include an *mm* data collection camera network, a *mm* keyframe extraction algorithm, and a *mm* time-series analysis algorithm to model variable pose duration and distinguish between sleep poses and transitory (or pseudo) poses. The views and modalities are shown in Figure 4.1 Block #1 with sample motion summaries shown in Block #6. The two resolutions are based on two of the most common ICU conditions: sleep hygiene and DU analysis. Pose history summarization is the coarser resolution. It provides a pictorial representation of poses over time. The applications of the pose history include prevention and analysis of DUs and analysis of sleep-pose effects on quality of sleep. The pose transition summarization is the finer resolution. MASH looks at the pseudo-poses that occur while a patient transitions between two poses. Applications of pose transition summarization include analyzing and quantifying physical therapy and distressed sleep motion quantification and analysis.

The MASH system is composed of three nodes. They are battery powered, enclosed by aluminum cases, controlled by Raspberry Pi3 [72] ARM-computers running Ubuntu 16.04 (to record video using a Carmine RGB-D sensor), and synchronized using TCP/IP communication, which are shown in Figure 4.2.

Multiple Modalities (Multimodal). Multimodal studies use complementary modalities to classify static sleep poses in natural ICU scenes with large variations in illumination and occlusions. MASH leverages the findings from [75] and [61] as evidence of the benefits of multimodal systems. The RGB-D views are shown in Figure 4.1 Block #1.

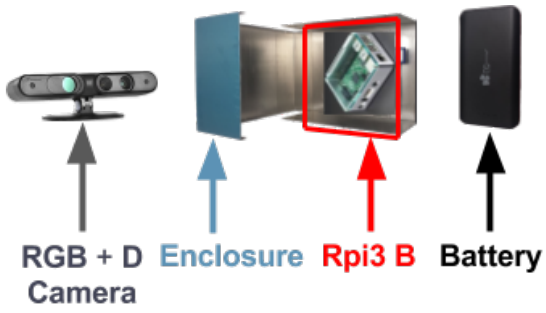


Figure 4.2: Elements of one MASH node.
From left-to-right: Carmine RGB-Depth sensor, aluminum enclosure, Raspberry Pi3 B+ with plastic enclosure and heat-sinks, and 24000 mAh battery. The elements are used to deploy MASH in the mock-up and the medical ICU rooms.

RGB (R). Standard video data provides information to represent and classify human sleep poses in scenes with relatively ideal conditions. However, most people sleep in imperfectly illuminated scenarios, using sheets, blankets, and pillows that block and disturb sensor measurements. The system collects RGB frames of dimensions 640×480 pixels. Pose appearance features representing human body configurations are extracted from these videos in BC and DO scenes.

Depth (D). Infrared depth cameras are resilient to illumination changes. The MASH sensor network uses Primesense's Carmine devices to collect depth data. The devices acquire images of dimensions 640×480 and use 16 bits to represent pixel intensity values, which correspond to the distance from the sensor to a point in the scene. Their operating distance range is 0.8 m to 3.5 m; and their spatial resolution for scenes 2.0 m away is 3.5 mm for the horizontal (x) and vertical (y) axes, and 30 mm along the depth (z) axis. The system uses the depth images to represent the 3-dimensional shape of the poses. However, depth information alone is not sufficient since it requires depth contrast, which is negatively affected by the deformation properties of mattresses, pillows, and blankets in the ICU.

Multiple Views (Multiview). The studies from [75] and [61] show that analyzing actions from multiple views and multiple orientations greatly improves detection. These studies indicate that the analysis of multiple views yield algorithms, which are independent of view and orientation. The positions of the cameras in the medical ICU are shown in Figures 4.4.

Time Analysis. ICU patients move subtly and slowly, very different from active motions like jumping or walking, which are easier to detect. MASH effectively monitors subtle and abrupt patient motion by breaking the motion cues into segments to flexibly model pose and pseudo-pose duration. The variable pose duration is modeled via HSMM, which is derived from conventional HMM.

Motion Quantization. The optic flow estimation is computed using the OpenCV [30] implementations of Lucas-Kanade [46] and Farneback [15]. Implementation and experimental results indicate that Lucas-Kanade led to faster results, while Farneback's led to higher accuracy in the detection of the most subtle pose transitions. Such pose transition is observed when transitioning from the left-yearner to the left-log positions without rotating. The two poses and their transition are shown on the bottom row of Figure 4.6 in green. From left to right, the second and third pose are yearner-left and log-left.

Inception CNN Feature Extraction. Deep feature extraction of using Google's Inception architecture required sizing the frames the appropriate image dimensions of 224×224 pixels. The offline analysis and approach uses Inception features due to the infrastructure restrictions, which prohibit the use large computation equipment. The deployed RPi3-based system cannot compute Inception features. Instead, the deployed system uses the online feature extraction method from [75].

4.3 The MASH Dataset

The views of the mock-up ICU are shown in Figure 4.1 Block #1 and the views of the medical ICU are show Figure 4.4. The fully annotated dataset will be available online to researchers. The real patient data is not controlled and only annotated after the fact. Table 4.3 shows the observed counts of poses in number of minutes. Figure 4.5 shows the count of pose transitions observed in the medical ICU. The cell colors indicate the transition is not applicable (N/A), had no rotation (gray), or had a left (orange), or right (green) rotation.

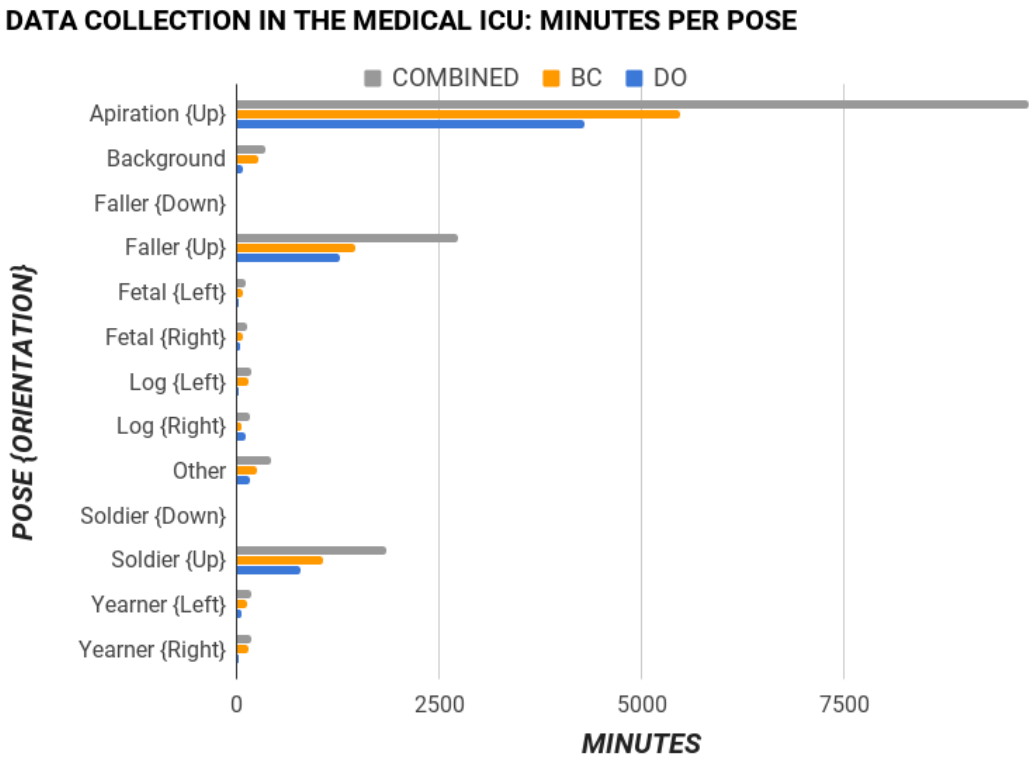


Figure 4.3: Number of minutes for poses recorded in the medical ICU.

The mock-up ICU. The room allows researchers to collect static and dynamic data, design and test algorithms, and evaluate and refine MASH.

Poses Static Data. The mock-up sequence is set at random. All actors in the mock-up ICU are asked to assume and hold each of the poses while videos are recorded. The combination of two separate recording sessions of six actors (three female and three male) yield a total of 24 sessions: 12 for BC and 12 for DO scene conditions. Each pose is recorded for one minute, which makes each session 10 minutes long.

Pose Transitions Data. The actors start in the initial pose and transition towards a final pose by rotating left or right. This processes is repeated for all initial poses and until all possible combinations between initial and final poses are covered. The combination of ten poses, with two possible transition rotations each generates a set of 20 sequences for each initial pose. Each recording session includes ten initial poses and ten final poses; therefore, each transition recording session generates 200 sequence pairs. A sample transition sequence with left and right rotation directions is shown in Figure 4.6. The initial and final poses are Faller Up ($falU$) and Fetal Left ($fetL$), respectively. The top sequence (orange) shows the left rotation and the bottom sequence shows the right (green) rotation. A small ($\leq 180^o$) rotation or a large ($> 180^o$) rotation are the possible transitions between the poses.

The medical ICU. The battery operated MASH network is currently deployed in a local community hospital where it is used to collect ICU data. The ICU patient dataset is thoroughly anonymized to protect the privacy of patients and medical staff. The dataset includes the video recordings of five consenting patients from periods of time that range from one to five days.

MASH Feature Extraction and Validation. Camera-based sleep pose classification studies commonly use geometric moments (gMOMs) and histograms of oriented gradients (HOG) to represent poses. Features extracted from video frames using convolutional

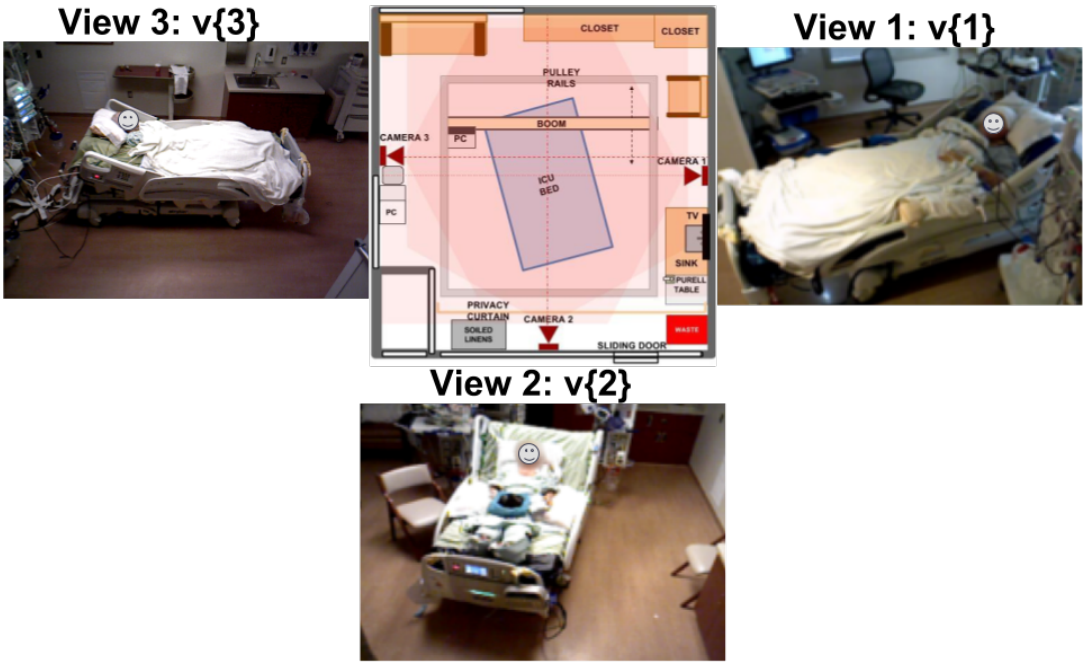


Figure 4.4: MASH node locations and views of the patient in the medical ICU.

IN PLACE: NO ROTATION	FINAL POSES																									
LEFT ROTATION	Aspiration		Faller U		Faller D		Fetal L		Fetal R		Log L		Log R		OTHER	Soldier D		Soldier U		Yearner L		Yearner R	
RIGHT ROTATION	LEFT	RIGHT	LEFT	RIGHT	LEFT	RIGHT	LEFT	RIGHT	LEFT	RIGHT	LEFT	RIGHT	LEFT	RIGHT		LEFT	RIGHT	LEFT	RIGHT	LEFT	RIGHT	LEFT	RIGHT			
INITIAL POSE: Apiration	N/A		8		N/A		2	N/A	N/A	3	2	N/A	N/A	3	2	N/A		8		5	N/A	N/A	4			
Faller D	N/A						N/A											N/A								
Faller U	7		N/A	N/A			4	N/A	N/A	3	3	N/A	N/A	2	4			N/A	N/A	0	N/A	N/A	0			
Fetal L	N/A	3	N/A	4			N/A		N/A	2	4		N/A	0	0			N/A	2	3		N/A	0			
Fetal R	2	N/A	3	N/A			1	N/A	N/A		0	N/A	3		1			0	N/A	0	N/A	5				
Log L	N/A	5	N/A	5			3		N/A	1	N/A		N/A	0	2			N/A	3	4		N/A	0			
Log R	2	N/A	4	N/A			3	N/A	2		1	N/A	N/A		0			1	N/A	1	N/A	3				
OTHER	2	4	2	3			N/A								3			3	3	N/A						
Soldier D	N/A						N/A											N/A								
Soldier U	6		4					N/A	N/A	2	4	N/A	N/A	2	3			N/A	N/A	0	N/A	N/A	0			
Yearner L	N/A	4	N/A	2			6	N/A	0		5		N/A	1	0			N/A	0	N/A	N/A	N/A	0			
Yearner R	3	N/A	1	N/A				N/A	5		0	N/A	5		0			0	N/A	0	N/A	N/A	N/A			

Figure 4.5: Pose transition count from the medical ICU recorded by MASH. The cell colors indicate the transition is not applicable (labeled N/A), the transition has no rotation (gray), left rotation (orange), or right rotation (green).

networks such as VGG [67] and Inception [71] architectures improve the pose classification performance of MASH. Pose classification results (see Section 4.5.1) indicate that using Inception outperforms using gMOMs, HOG, and VGG features. Feature extraction of

gMOM and HOG features is based on the parameters from [77]. The methods from [24] are used to calibrate the cameras and subtract the background, prior to feature extraction.

4.4 The MASH Approach

In order to effectively analyze patient motion, the MASH needs to handle variable motion rates (speed) and motion ranges (rotation angle) simultaneously. The initial assumption for all video frames is that they belong to pose transitions (pseudo-poses), but if the motion rate is identified as slow, these frames can be used to identify true poses, which are needed to identify pose histories (i.e., the sequence of poses). The pose transition involves identifying the set of pseudo poses representing a transition between two poses, and it quantifies the direction of rotation. The first challenge arises because conventional algorithms are unable to model pose duration effectively. The second challenges involves detecting the direction of rotation when transitioning between poses. The last challenge involves representing pseudo poses, for which MASH uses keyframe estimation. The M multimodal cameras are stationed at different locations to obtaining V views of the patients as shown in Figure 4.4 and estimate the pose transition dynamics, such as the ones in Figure 4.6.

The features extracted from video frames $\mathcal{F} = \{f_t\}$, for $1 \leq t \leq T$ to construct feature vectors $\mathbf{X} = X_{1:T}$ are used to represent non-directly observable poses ($\mathbf{Y} = Y_{1:T}$). The first objective of MASH is to find the sequence of poses ($\mathbf{Y} = Y_{1:T}$) that probabilistically can best represent the observations, as in: $\Pr\left(\mathbf{Y}, \mathbf{X}\right) = \Pr\left(Y_{1:T}, X_{1:T}\right)$. Temporal patterns caused by sleep-pose transitions are simulated and analyzed using Hidden Semi-Markov Modeling (HSMMs) technique, which is described in Section 4.4.2. The interactions between the modalities for accurate pose representation are encoded into the emission probabilities. Scene conditions are encoded into the set of states (the

analysis of two scenes doubles the number of poses). Conventional Markov assumptions support MASH and ideally fit most of its analysis. However, HMMS are limited in their ability to distinguish between poses and pseudo-poses based on pose duration. This is because, by design, HMMs model the probability of staying in a given pose as a geometric distribution $\Pr_i(d) = (a_{ii})^{d-1}(1 - a_{ii})$, where d is the duration in pose i, and a_{ii} is the self-transition probability of pose i. More details are discussed in Sections 4.4.1 and 4.4.2. Table 4.1 describes the MASH variables.

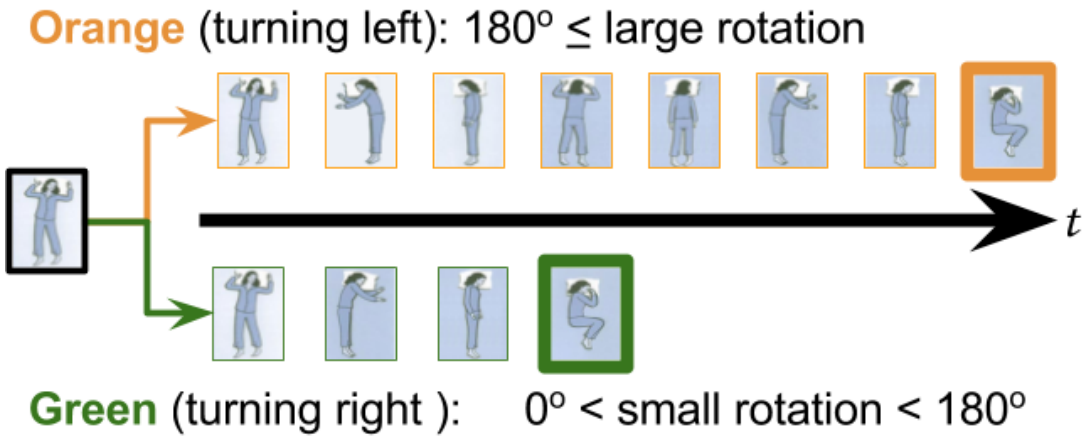


Figure 4.6: Two ways to rotate between transitions.
Pose transitions require patients to reconfigure their body. In this example, transitioning from the faller facing up ($falU$) position to the fetal laying on the left ($fetL$) position can be achieved by either a long rotation (180^o; top row) or by a short rotation ($0 - 180^o$; bottom row).

MASH VARIABLES

SYMBOL	DESCRIPTION
$\mathbf{A}$	Transition probability matrix $\mathbf{A} \in \mathbb{R}^{\lvert P\rvert \times \lvert P\rvert}$ and $\mathbf{A} = \{a_{ij}\}$
$a_{i,j}$	Probability of transition from pose i to j
$\mathbf{B}$	Emission probability matrix $\in \mathbb{R}^{\lvert P\rvert}$ and $\mathbf{B} = \{\mu_{in}\}$
b_u	Beginning of the u-th segment with $b_1 = 1$
D_k	k-th frame from the depth modality video
D	Face-Down patient pose
d	Segment duration
d_u	Segment duration for u-th segment
HMM	Abbreviation for Hidden Markov Model
HSMM	Abbreviation for Hidden Semi-Markov Model
K	Data set size, $K = \lvert\mathcal{X}\rvert$
k	Data point index, $1 \leq k \leq K$
KF	Set of sequential keyframes representing a transition between
L	Laying-Left patient pose

l, m, and n	Dummy variables
R_k	k-th frame from the rgb modality video
R	Laying-Right patient pose
μ_i	Probability that state i generates the observation x at time t
π	Initial state probability vector $\in \mathbb{R}^{\lvert}P\rvert$ and $\pi_i \in \pi$
k	The time step index (i.e., $k = t$)
P	Set of patient poses $P = \{p_i\}$
P_{mock}	Set of actor poses in the mock-up ICU room
P_{micu}	Set of patient poses in the real medical ICU (micu) room
$\Pr(Y, X)$	The joint PDF: sequence of states and observations
t	Time tick with $1 \leq t \leq T$
τ_{td}	Stores the estimated duration ($1 \leq d \leq D$) at time (t)
θ	HMM model with probabilities $\mathbf{A}, \mathbf{B}$, and $\mathbf{\Pi}$
U	Number of segments $U = \lvert S\rvert$
U	Face-Up patient pose
u	Segment index: $1 \leq u \leq U$
$\mathcal{V}$	View set $\mathcal{V} = \{$left, center, right$\}$
V	Number of views $V = \lvert\mathcal{V}\rvert$
v	View index, $1 \leq v \leq V$
$\mathcal{X}$	Dataset indexed by k (i.e., $\mathcal{X}_k$)
$\mathcal{X}_k$	k-th datapoint with $\{f_{N_m}\}_k = \{f_R, f_D, f_P\}_k$
x_k	k-th observation feature vector
$\mathbf{x}_{km}^{(v)}$	The k-th observable variable from view v and modality m
y_k	k-th hidden state $y_k \in \mathbf{Y}$
$\mathbf{Y}$	Sequence of hidden states $\lvert\mathbf{Y}\rvert = T$
δ	Kroenecker delta function
δ_t	The maximum probability duration
$\mathbf{\Omega}$	Set of time segments $\{\Omega_u\}$ for $1 \leq u \leq U$
Ω	Segment element $\Omega \in \mathbf{\Omega}$
θ	Dummy variable used in inference
ζ	Stores the state label (for a pose) of the previous segment
ϕ	Stores the best duration
$\psi_t(i)$	Stores the label with the best duration for state i at t

Table 4.1: MASH variables and their descriptions.

4.4.1 Hidden Markov Models (HMMs).

HMM is a generative modeling approach that represents pose history and transitions as states. The hidden variable or state at time step k (i.e., $t = k$) is y_k (state$_k$ or pose$_k$) and the observable or measurable variables ($x_{k,m}^{(v)}$, the vector of image features corresponding to the k-th frame, using modality m, and view v) at time $t = k$: x_k such that $x_k = x_{k,m}^{(v)} = \{R_k, D_k, \ldots M_k\}$). The Markovian assumptions indicate that at t, the hidden variable y_t, depends only on the previous hidden variable y_{t-1}, and at t the observable variable x_t depends on the hidden variable y_t. These two assumptions are used to compute $\Pr(Y, X)$ given by:

$$\Pr\left(Y_{1:T}, X_{1:T}\right) = \Pr(y_1) \prod_{t=1}^{T} \Pr\left(x_t|y_t\right) \prod_{t=2}^{T} \Pr\left(y_t|y_{t-1}\right), \tag{4.1}$$

where $\Pr(y_1)$ is the initial state probability distribution ($\mathbf{\Pi}$). It represents the probability of a sequence starting at ($t = 1$) pose$_i$ (state$_i$). $\Pr\left(x_t|y_t\right)$ is the emission probability distribution ($\mathbf{B}$) and represents the probability that at time t, y_i (state$_i$) can generate the observable multimodal multiview vector x_t. Finally, $\Pr\left(y_t|y_{t-1}\right)$ is the transition probability distribution ($\mathbf{A}$) and represents the probability of going from pose$_i$ to pose$_o$ (state i to o). The standard HMM parameters are: $\mathbf{A} = \{a_{ij}\}$, $\mathbf{B} = \{\mu_{in}\}$, and $\mathbf{\Pi} = \{\pi_i\}$.

Modeling Limitations of HMM. One critical limitation of HMM is its rigidity to model state duration. For instance, given an HMM in a state i (pose or transition), the probability that it stays there for d time slices is: $\Pr_i(d) = (a_{ii})^{d-1}(1 - a_{ii})$, where $\Pr_i(d)$ is the discrete probability density function (PDF) of duration d in pose i, and a_{ii} is the self-transition probability of pose i, given by a geometric distribution [60]. The inability to flexibly model pose and transition duration is observed when similar body positions can only be discerned by their distinctive duration (pose vs transitory pose). This limitation is tackled using HSMM and is described in section 4.4.2.

4.4.2 Hidden Semi-Markov Models (HSMMs)

HSMM serves to flexibly model state duration. It uses segments instead of time slices to sample observations. In HSMM, hidden variables are segments, which have useful properties. Figure 4.1 Block #5 shows the HSMM trellis and indicates its main components. For instance, the sequence of states $y_{1:T}$ is represented by the segments ($\mathbf{\Omega}$). A segment is a sequence of unique, sequentially repeated poses (symbols), which serves to identify and track an observation's first instance and the observation's duration (based on the number

of observed samples). From the original sequence, the elements of the j-th segment (Ω_j) are the indices at which the observation (b_j) is first detected; the number of sequential observations of the same symbol (d_j); and the state or pose symbol (y_j). For instance, the sequence $y_{1:9} = \{4, 4, 2, 2, 2, 3, 2, 1\}$ is represented by the set of segments $\mathbf{\Omega} = \Omega_{1:U}$ with elements $\Omega_{1:U} = \{\Omega_1, \Omega_2, \Omega_3, \Omega_4, \Omega_5\} = \{(1, 2, 4),\ (3, 3, 2),\ (6, 1, 3),\ (7, 1, 2),\ (8, 1, 1)\}$, where U is the total number of segments (i.e., state changes). The elements of the segment $\Omega_{j=1} = (b = 1, d = 2, y = 4)$ indicate that the segment started at the first observation, lasted two time samples, and was observed to be the fourth state.

HSMM components. In conventional HMM, the hidden variables are y, but in HSMM, the hidden variables are now the segments $\mathbf{\Omega} = \Omega_{1:U}$, while the observable features are the same in both methods ($X_{1:T}$). The joint probability of the segments $\Omega_{1:U}$ and the observable variable $X_{1:T}$ is:

$$\begin{aligned}
\Pr\left(\Omega_{1:U}, X_{1:T}\right) &= \Pr\left(Y_{1:U}, b_{1:U}, d_{1:U}, X_{1:T}\right) \\
\Pr\left(\Omega_{1:U}, X_{1:T}\right) &= \Pr(y_1)\Pr(b_1)\Pr(d_1|y_1)\prod_{t=b_1}^{b_1+d_1+1}\Pr(x_t|y_1)\times \\
&\quad \prod_{u=2}^{U}\Pr(y_u|y_{u-1})\Pr\left(b_u|b_{u-1}, d_{u-1}\right)\times \\
&\quad \Pr\left(d_u|y_u\right)\prod_{t=b_u}^{b_u+d_u+1}\Pr(x_t|y_u).
\end{aligned} \tag{4.2}$$

Recall that U is the sequence of segments such that $\Omega_{1:U} = \{\Omega_1, \ldots, \Omega_U\}$ for $\Omega_u = \left(b_u, d_u, y_u\right)$, b_u as the start position (a bookkeeping variable to track the starting point of a segment), d_u is the duration, and y_u is the hidden state ($\in \{1, ..., Q\}$). The range of time slices starting at b_u and ending at $b_u + d_u$ (exclusively) have state label y_u. All segments have a positive duration and over the time-span $1 : T$ without overlap and with

constraints $b_1 = 1$, $\sum_{u=1}^{U} = T$ and $b_{u+1} = b_u + d_u$.

The transition probability $\Pr(y_u|y_{u-1})$, is the probability of going from one segment to the next via:

$$\mathbf{A} : \Pr\big(y_u = j|y_{u-1} = i\big) \equiv a_{ij} \tag{4.3}$$

The first segment (b_u) starts at 1 ($u = 1$) and consecutive points are calculated from the previous point using the following:

$$\Pr\big(b_u = m|b_{u-1} = n, d_{u-1} = l\big) = \delta\big(m - n - l\big) \tag{4.4}$$

where $\delta(i - j)$ is the Kroenecker function with value of 1, if $i = j$ and 0, else (i.e., $m = n + l$). The duration probability is now given by $\Pr(d_u = l|y_u = i) = \Pr_i(l)$, where $\Pr_i(l)$ is now a free parameter. This allows MASH to sample a distribution of the form $\Pr_i(l) = \mathcal{N}(\mu, \sigma)$ in the software implementation. A normal distribution serves to compute the duration probability of the i-th state and distinguishing between slow and fast pose duration/transitions. This section also covers the estimation of MASH parameters, Viterbi computation, and inference.

MASH Parameter Estimation. HSMM estimation of parameters is based on maximum likelihood (MLE). The training sequence of keyframes is fully annotated, including the start and end index frames for each segment $X_{1:T}, Y_{1:T}$. To find the parameters that maximize $\Pr\big(Y_{1:T}, X_{1:T}|\theta\big)$, the likelihood parameters of each of the factors in the joint probability must be maximized. In particular, the observation probability, $\Pr\big(x^n|y = i\big)$,

is a Bernoulli distribution whose maximum likelihood is computed as follows:

$$\mu_{n,i} = \frac{\sum_{n=1}^{T} x_n^i \delta\big(y_n, i\big)}{\sum_{n=1}^{T} \delta\big(y_n, i\big)}, \tag{4.5}$$

where T is the number of time-series data points, $\delta(i, j)$ is the Kroenecker delta function, and $\Pr\big(y_t = j | y_{t-1} = i\big)$ is the multinomial distribution of the form:

$$a_{ij} = \frac{\sum_{n=2}^{N} \delta\big(y_n, j\big)\delta\big(y_{n-1}, i\big)}{\sum_{n=2}^{N} \delta\big(y_{n-1}, j\big)} \tag{4.6}$$

Viterbi for MASH. The segment notation is used to represent state sequences in HSMM modeling. The objective behind the inference is to find the state sequence that maximizes $P\big(\Omega_{1:U}, X_{1:T}|\theta\big)$, for a new sequence of observations with unknown duration. The sequence corresponding to the duration with the highest probability is determined at each time step by iterating over all possible duration values from 1 to a predetermined duration D. This data is stored in:

$$\tau_{t,d,i} = \max_{\Omega_1,\ldots,\Omega_{k-1}} \Pr\Big(X_{1:t}, \Omega_{1:k} = \big(t - d + 1, d, i\big)|\theta\Big), \tag{4.7}$$

which represents the highest probability of a sequence of K segments, where the final segment started at $t - d + 1$, has duration d and label i. In conventional HMMs, it is sufficient to only keep track of the maximum probability of ending in state Ω_{k-1} to compute the maximum probability of ending up in state Ω_k.

The label for a pose or state of the previous segment is stored in $\zeta_t(d, i)$. The maximum probability duration (δ) is computed via:

$$\delta_t(i) = \max_{\Omega_1,\ldots,\Omega_{k-1}} \Pr\Big(x_{1:t}, \Omega_{1:k} = \big(t - d^* + 1, d^*, i\big)|\theta\Big), \tag{4.8}$$

where d^* is the duration with the highest probability at time t for state i. The variables $\phi_t(i)$ and $\psi_t(i)$ store the best duration and the label of the previous segment, respectively.

Inference for MASH. Four steps for finding the best sequence:

1. *Initialization.* The label probability of the first segment comes from the initial state distribution π and computed via

$$\tau_{t,d} = \pi_i \Pr_i(d) \prod_{t=1}^{T} \Pr\left(x_t | y_t\right) \text{ and } \zeta_d(d,i) = 0. \tag{4.9}$$

2. *Recursion.* Iterate over all possible duration values in given by:

$$\tau_{t,d} = \max_{1 \le i \le Q} \left[\delta_{t-d}(i) a_{ij}\right] \Pr_j(d) \prod_{m=m_1}^{t} \Pr\left(\vec{x}_m | y_m = j\right), \tag{4.10}$$

   where $m_1 = t - d + 1$ and $\zeta_d(d,i) = \arg\max_{1 \le i \le Q} \left[\delta_{t-d}(i) a_{ij}\right]$.

   The duration with the highest probability is $\delta_t(i) = \max_{1 \le d \le D} \left[\delta_{t-d}(i) a_{ij}\right]$, which represents the best segment. The variable d^* is the duration with the highest probability at time t for state i. The best duration for state i at time t is computed from $\phi_t(i) = \arg\max_{1 \le d \le D} \tau_{d,t}(i)$. Finally, $\psi_t(i) = \zeta_t\big(\phi_t(i), i\big)$ represents the label of the best duration at time t for state i.

3. *Termination.* Estimate the state with the highest probability in the last slice via $\Pr^* = \max_{1 \le i \le Q} [\delta_T(i)]$, where $y_T^* = \arg\max[\delta_T(i)]$, $t = T$, and $u = 0$.

4. *Backtracking.* From the termination, look up the duration and previous states stored in variables ϕ and ψ given by $d_t^* = \phi_t\big(y_t^*\big)]$ and $\Omega_u^* = \big(t - d_t^* + 1, d_t^*, y_t^*\big)$, with $t = t - d_t^*$, $u = u - 1$, and $y_t^* = \phi_{t+d}(y_{t+d}^*)$.

Negative indexing is used for the segments since their number is unknown. This is corrected after the inference step by simply adding $|\mathbf{\Omega}^*|$ to all indices.

Keyframe (KF) Selection. Pose datasets are very large, often repetitive, and with relatively slow and subtle motion. The pre-processing stage uses a keyframe estimation applied to mm data. The extraction algorithm used to extract the set (KF) of K-transitory frames is shown in Figure 4.7 and detailed in Algorithm 2. The size of the set is determined experimentally ($K = 5$) on the feature space using Inception vectors.

Algorithm 2 Multimodal multiview keyframe selection using euclidean dissimilarity measure. The algorithm is applied at training with labeled frames to estimate the number and indices of keyframes across views and modalities.

procedure -Inputs: ($\mathcal{X}$, set of mm features and dissimilarity threshold th)
  $KF = \{\text{Keyframes}\}_K$, $K \geq 1$
  Initialize: $KF = \{\text{empty}\}_K$, $K \geq 1$ and $count = 0$
  procedure -stage 1: (Modality (m) and View (v) Selection)
    for $1 < v < V$ and $1 < m < M$ **do**
      $D_m^{(v)} = \text{euclid}(x_{mn_i}^{(v)}, x_{mn_o}^{(v)})$, $n_i = 1, n_o = N$
      $\hat{v}, \hat{m} = \max D_m^{(v)} > th$
      $KF \leftarrow \{x_{\hat{m}n_1}^{(\hat{v})}, x_{\hat{m}n_N}^{(\hat{v})}\}$
    end for
  end procedure
  procedure -stage 2: (Find Complementary Frames to KF)
    for $1 < v < V$ and $1 < m < M$ and $1 < n < N$ **do**
      $D_1 = D_{m,n_1}^{(v)} = \text{euclid}(x_{mn_1}^{(v)}, x_{mn}^{(v)})$
      $D_2 = D_{m,n_N}^{(v)} = \text{euclid}(x_{mn_N}^{(v)}, x_{mn}^{(v)})$
    end for
    Sort $D_1, D_2 = \{d_1 > d_2 > ... > d_{N-2}\}$, descending
    $KF \leftarrow d_i$ if $\frac{d_i}{d_j} > th$, for $1 < i, j < N - 2$
  end procedure
  procedure -stage 3: (Find Center Frame (i.e., Motion Peak))
    for KF_2 and KF_{K-1} **do** Use Stage 2 to compute D_3 and D_4
      if $\max(D_3, D_4) > 0$) **then** $\max(D_3, D_4) \rightarrow KF$
      end if
    end for
  end procedure
end procedure

Let $\mathcal{X} = \{x_{m,n}^{(v)}\}_f$ be the set of training features extracted from V views and M modalities over N frames and let P_i and P_o represent the initial and final poses. The transition frames are indexed by n, $1 \leq n \leq |N|$; views are indexed by v, $1 \leq v \leq |V|$ and modalities are indexed by m, $1 \leq m \leq |\mathcal{M}|$. Algorithm 2 uses this information to

identify keyframes. Experimental evaluation of $|KF|$ is shown in Figure 4.5.2. Keyframes are the most informative and discriminant frames for all views and modalities.

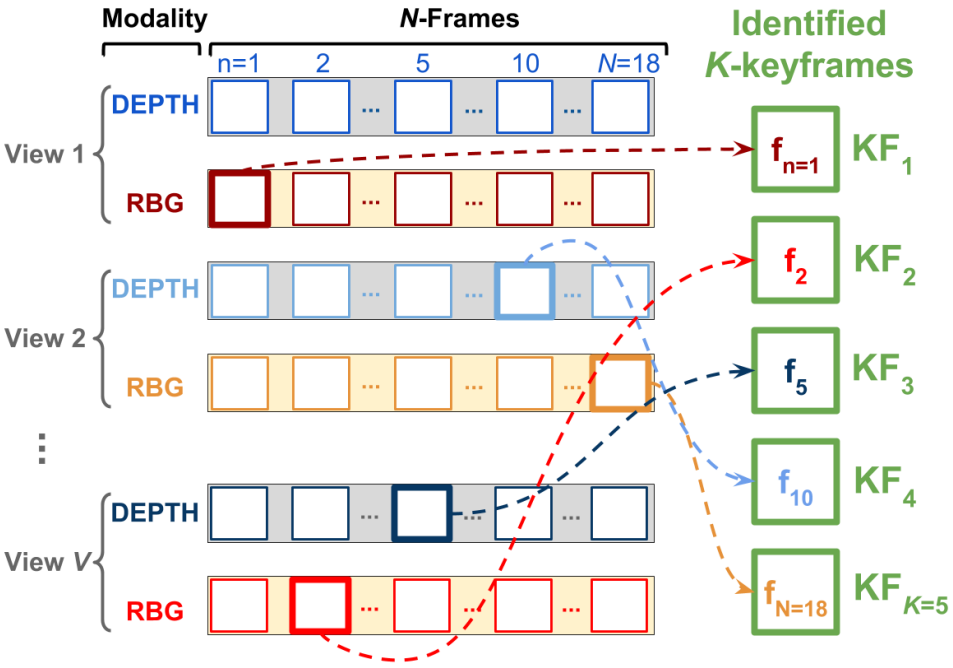


Figure 4.7: Keyframe extraction for pose transition representation.
The keyframe selection is based on Algorithm 2. This figure shows MASH's keyframe extraction process from three views and two modalities. The first two keyframes are extracted from the first camera's RGB modality (Views 1 and 2). Subsequent keyframes are selected from View 2's depth, and View V's RGB.

4.5 MASH Experimental Results

MASH is evaluated using a five-fold cross-validation approach. The results indicate that deep features increase MASH's classification accuracy over engineered features by 7% in DO scenes (from 86.7% to 93.6%), while matching the performance of engineered features in BC scenes. The overall time tracing and summarization error rate between HMM and the proposed MASH approach increased from 46.4% to 83.2% in the mock-up

Feature Suitability Evaluation with cc-LS [75]			
Scene	HOG + gMOM	Vgg	Inception
BC	100	100	100
DO	86.7	90 (+3)	93.6 (+7)

Table 4.2: Evaluation of features for sleep-pose recognition.
The evaluation uses the cc-LS method from [75] in dark and occluded scenes. The performance of HOG and gMOM features is outperformed by the performance of VGG and Inception features.

ICU and from 35.8% to 80.1% in the medical ICU. In addition, the proposed keyframe transition representation achieves a classification of 78%.

4.5.1 Static Pose Analysis - Feature Validation.

Static sleep-pose classification analysis is used to compare the MASH method to previous studies. Couple-Constrained Least-Squares (cc-LS) [75] and MASH are tested on the dataset from [75]. Combining the cc-LS method with deep features extracted from two common network architectures improved classification performance over the HOG and gMOM features in DO scenes by an average of eight percent with Inception and four percent with VGG. Deep features matched the performance of cc-LS (with HOG and gMOM) for a BC scenario. Results for both scenes are shown in Table 4.5.1. Similarly, the contribution of each of the multimodal and multiview sources is analyzed and evaluated. The plot in Fig. 4.8 shows the contribution of each MASH sensor modality and view to the mean classification accuracy of static poses using cc-LS from [75].

Similarly, the contribution of each of the multimodal and multiview sources is analyzed and evaluated. Figure 4.8 shows the contribution of each MASH sensor modality and view to the mean classification accuracy of static poses using cc-LS.

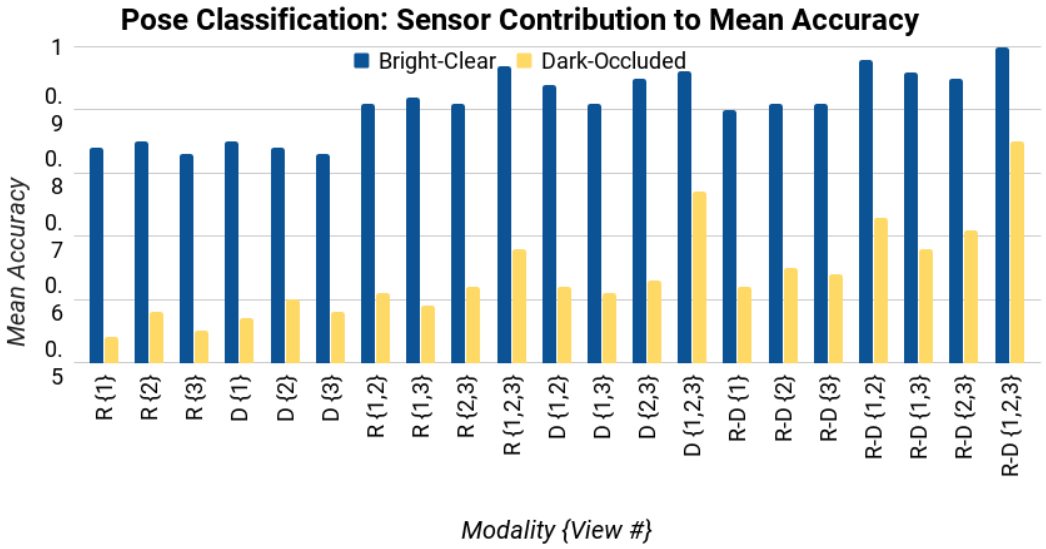


Figure 4.8: Sensor contribution to mean classification accuracy of sleep poses. MASH performance is evaluated in BC (yellow) and DO (blue) scenes. The RGB and Depth modalities are represented by R and D, respectively. The camera views shown in Figure 4.4 are marked using {view number}.

4.5.2 Keyframe Performance.

The effect of $|KF|$ $(= 5)$ and keyframe dissimilarity threshold th $(\geq .8)$ on pose transition classification accuracy is shown in Figure 4.5.2. The traces indicate transitions correctly identified by MASH.

4.5.3 Summarization Performance

Pose history summarization is important to decubitus ulceration prevention and analysis. An example of the objective behind history summarization is shown in Figure 4.10, where the sequence of poses is identified as A or B. History summarization is the coarser time resolution. The mock-up ICU enables staging the motion and scene condition variations necessary to carry out this experiment. In particular, it avoids disturbing real

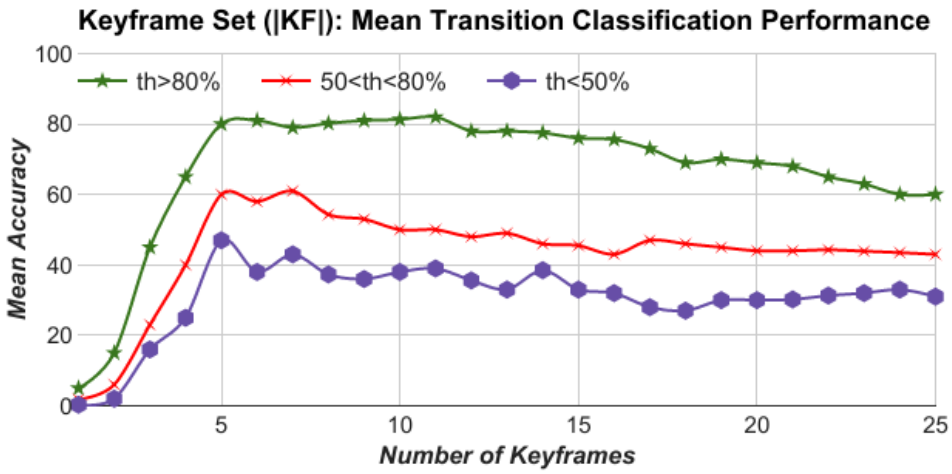


Figure 4.9: Motion summarization dependency on keyframe set size.
The number of keyframes used to represent transitions and rotations between poses have a direct impact on summarization. The best performing set has size 5 and dissimilarity threshold (th) value of 0.80.

patients in the medical ICU. Table 4.5.3 contains the numerical symbols of the various poses and the names used in the summarization traces.

MASH: Pose History Summarization

Symbol	Pose Name
0	Aspiration
+1 / -1	Soldier (+Up / -Down)
+2 / -2	Yearner (+R / -L)
+3 / -3	Log (+R, -L)
+4 / -4	Faller (+Up / -Down)
+5 / -5	Other / Background
+6 / -6	Fetal (+R / -L)

Table 4.3: Pose symbols and descriptions.
The symbols are used for ICU pose history summarization in the mock-up and the medical rooms.

Pose History Summarization in the ICU

Summarization history results are shown in Figure 4.11 for the mock-up ICU and Figure 4.12 for the medical ICU room in (b). The accuracy is computed as the percent

overlap between the trace representing the true poses and the traces representing MASH and HMM in orange and gray, respectively. The pose history summarization experiments are staged using a sampling rate of one second and an pose duration of 10 seconds, with a minimum average detection of 80 percent. A pose is assigned a label if it is consistently detected (i.e., 80% of the time), including the label "other". Poses that are not consistently detected are ignored. The system is tested in the mock-up setting using a randomly selected sequence of ten poses starting with a randomly selected scene condition. The duration of the poses is also selected at random with one scene transition (from BC to DO or from DO to BC). The history summarization performance is shown in Table 4.4.

MASH: Pose History Summarization	
Scene	Average Detection Rate
BC	85
DO	76

Table 4.4: MASH pose history summarization performance.
The MASH framework in BC and DO scenes in the mock-up ICU. The sequences are composed of 10 poses with duration ranging from 10 seconds to 1 minute and with sampling rate of one second.

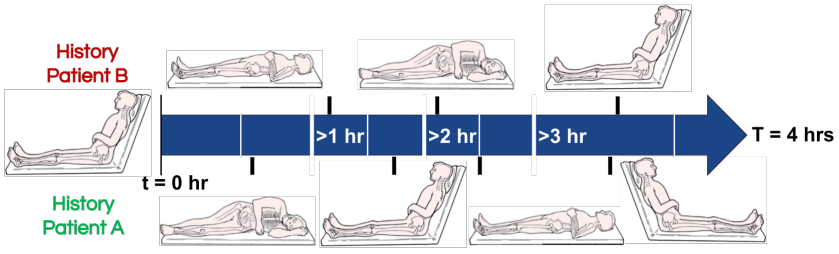


Figure 4.10: Sample pose history summarization log.
The sample pose summary covers a 4-hour span.

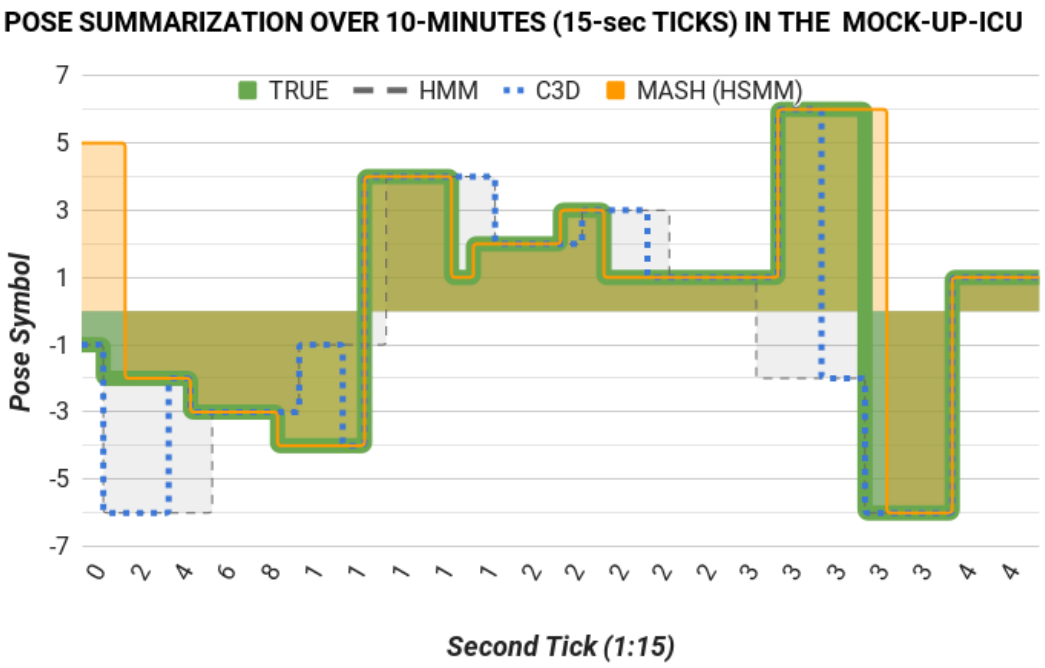


Figure 4.11: MASH mock-up ICU history summarization.
The traces are the ground-truth (solid green) poses and HMM (dashed gray), C3D (dotted blue), and MASH (solid orange) pose summaries under BC conditions on a 10-minute video.

4.5.4 Pose Transition Dynamics: Motion Direction.

The detection and quantization of transitions and directions of rotations is important to physical therapy and recovery rate analysis. The contribution of MASH sensors and views to pose transition classification accuracy is shown in Figure 4.15.

Transition Summarization in the Mock-Up ICU

The performance of MASH summarizing fine motion to describe transitions between poses is shown in Figures 4.13 and 4.14 for (a) singleview and (b) multiview system configurations, while (c) shows the legend. Similarly to pose classification, the multi-modal and multiview elements in MASH are complementary to pose transition analysis.

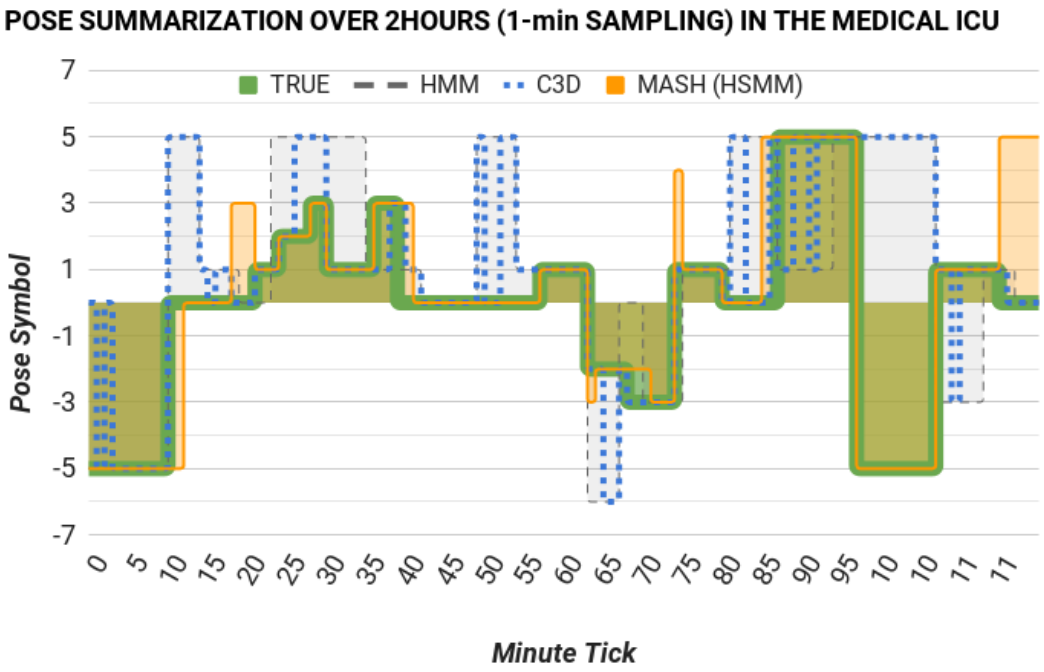


Figure 4.12: MASH medical ICU history summarization.
The traces are the ground-truth (solid green) poses and HMM (dashed gray), C3D (dotted blue), and MASH (solid orange) pose summaries under random scene conditions on a two-hour video with a reduced set of poses due to patient immobility. The medical summary is based on a two-hour medical round standard.

MASH's peak performance is attributed to the combination of multiple views and modalities. The contributions of each sensor and view to pose transitions classification accuracy are shown in Figure 4.15.

Transition Summarization in the Medical ICU.

Note that it is logistically impossible to control ICU work flows and to account for unpredictable patient motion in a medical ICU. ICU patients do not have the same rotation range as the patients/actors in the mock-up ICU. This mobility constraint reduces the set of poses and pose transitions (unavailable transitions are marked N/A). The timeline

REAL ICU		FINAL POSE: Po -- SINGLEVIEW									
BC SCENE	ROTATION	solU	solD	logR	logL	yeaR	yeaL	fetR	fetL	falU	falD
INITIAL POSE: Pi											
solU	LEFT <--	79	75	70	69	73	75	72	70	70	77
	RIGHT -->	76	72	72	68	75	70	76	69	69	79
	INPLACE	K								72	
solD	LEFT <--	74	77	75	73	77	71	76	71	74	73
	RIGHT -->	75	75	69	66	69	68	69	75	73	70
	INPLACE		K								76
logR	LEFT <--	67	71	67	71	70	75	71	63	76	68
	RIGHT -->	64	63	71	70	68	74	65	75	62	75
	INPLACE			K		71		73			
logL	LEFT <--	60	66	68	62	75	70	73	70	68	79
	RIGHT -->	73	76	76	66	70	71	70	70	73	65
	INPLACE				K		72		76		
yeaR	LEFT <--	74	63	73	66	74	73	78	76	76	76
	RIGHT -->	65	78	71	64	71	69	76	77	67	73
	INPLACE			76		K		81			
yeaL	LEFT <--	58	78	77	78	74	75	74	71	62	64
	RIGHT -->	70	63	72	74	73	71	73	79	73	66
	INPLACE				84		K		80		
fetR	LEFT <--	68	60	70	75	66	69	77	68	71	68
	RIGHT -->	62	74	75	75	70	66	78	75	66	74
	INPLACE			77		73		K			
fetL	LEFT <--	67	72	73	77	73	74	66	74	69	66
	RIGHT -->	75	63	81	75	64	66	70	77	75	73
	INPLACE				79		74		K		
falU	LEFT <--	68	63	65	72	66	73	76	68	76	71
	RIGHT -->	70	67	73	65	72	75	78	74	77	69
	INPLACE	75								K	
falD	LEFT <--	76	74	78	74	75	70	73	66	76	79
	RIGHT -->	74	72	61	61	77	74	65	74	79	77
	INPLACE		77								K

(a)

REAL ICU		FINAL POSE: Po -- MULTIVIEW									
BC SCENE	ROTATION	solU	solD	logR	logL	yeaR	yeaL	fetR	fetL	falU	falD
INITIAL POSE: Pi											
solU	LEFT <--	88	84	79	79	81	88	83	81	80	89
	RIGHT -->	86	84	85	73	R	87	88	77	81	91
	INPLACE	K								85	
solD	LEFT <--	83	87	84	84	86	82	89	79	89	81
	RIGHT -->	83	86	78	75	78	76	78	84	87	78
	INPLACE		K								85
logR	LEFT <--	87	80	78	83	82	84	78	76	88	76
	RIGHT -->	73	73	80	83	82	83	77	87	72	89
	INPLACE			K		86		81			
logL	LEFT <--	68	79	80	76	81	79	81	78	78	93
	RIGHT -->	84	84	85	78	80	80	83	81	83	72
	INPLACE				K		81		82		
yeaR	LEFT <--	84	78	84	79	83	84	89	87	86	88
	RIGHT -->	70	82	82	76	83	80	86	85	77	85
	INPLACE			85		K		90			
yeaL	LEFT <--	65	86	86	84	83	84	82	84	72	71
	RIGHT -->	84	76	84	82	84	84	86	86	83	74
	INPLACE				87		K		86		
fetR	LEFT <--	80	75	81	86	75	73	88	74	81	80
	RIGHT -->	71	86	82	86	80	75	88	87	76	83
	INPLACE			85		83		K			
fetL	LEFT <--	80	84	84	82	81	86	78	85	79	76
	RIGHT -->	84	71	89	80	73	78	79	85	85	85
	INPLACE				85		88		K		
falU	LEFT <--	75	74	76	83	78	88	86	89	86	80
	RIGHT -->	78	76	84	79	83	82	89	72	87	78
	INPLACE	81								K	
falD	LEFT <--	86	87	91	90	86	78	84	75	92	92
	RIGHT -->	84	87	74	75	85	82	71	86	89	92
	INPLACE		88								K

(b)

LEGEND					
COLOR	Back	Red	Geen	Bue	Orange
ROTATION ANGLE	0	90	180	270	360

(c)

Figure 4.13: MASH pose transition analysis in the BC mock-up ICU.
The analysis reports on detection and classification mean accuracy in the mock-up ICU under BC scene conditions for the singleview (a) and multiview (b) system configurations with the legend shown in (c).

in Figure 4.10 shows the clinical objective of pose history summarization. Once in production, the summarized pose history will be labeled by clinicians to indicate good/bad poses and correlate patterns to patient health status (i.e., replace the labels A and B with medically meaningful labels).

Views of the medical ICU room are shown in Figure 4.4 and the traced detections are shown in Figure 4.12. The green trace represents the true transition labels and the red

MOCKUP ICU		FINAL POSE: Po -- SINGLEVIEW									
DO SCENE	ROTATION	solU	solD	logR	logL	yeaR	yeaL	fetR	fetL	falU	falD
INITIAL POSE: Pi	solU LEFT <--	48	55	34	49	31	48	52	52	40	49
	solU RIGHT -->	43	48	55	43	47	49	48	48	51	61
	solU INPLACE	K								47	
	solD LEFT <--	43	37	54	54	46	42	39	52	36	41
	solD RIGHT -->	56	58	48	55	48	56	53	54	47	48
	solD INPLACE		K								45
	logR LEFT <--	47	50	48	33	22	48	68	67	48	53
	logR RIGHT -->	53	65	30	43	42	47	54	61	56	49
	logR INPLACE			K		45		61			
	logL LEFT <--	68	45	35	26	21	49	41	38	52	33
	logL RIGHT -->	54	36	49	48	34	55	53	41	63	52
	logL INPLACE				K		54		45		
	yeaR LEFT <--	54	58	44	69	36	34	44	58	52	48
	yeaR RIGHT -->	50	52	50	52	43	30	36	55	47	35
	yeaR INPLACE			52		K		50			
	yeaL LEFT <--	55	46	48	34	59	48	51	44	42	39
	yeaL RIGHT -->	57	37	55	41	63	56	55	47	53	34
	yeaL INPLACE				45		K		50		
	fetR LEFT <--	61	45	48	61	55	51	51	38	49	47
	fetR RIGHT -->	42	43	52	55	40	46	48	36	41	53
	fetR INPLACE			54		48		K			
	fetL LEFT <--	41	60	54	55	43	59	54	55	49	46
	fetL RIGHT -->	54	53	59	48	53	58	47	55	55	58
	fetL INPLACE				50		59		K		
	falU LEFT <--	35	33	27	61	68	75	56	66	32	44
	falU RIGHT -->	38	41	48	41	71	62	61	63	35	35
	falU INPLACE	40								K	
	falD LEFT <--	41	56	50	41	43	51	46	43	48	56
	falD RIGHT -->	43	54	50	50	40	53	45	45	46	43
	falD INPLACE		50								K

(a)

MOCKUP ICU		FINAL POSE: Po -- MULTIVIEW									
DO SCENE	ROTATION	solU	solD	logR	logL	yeaR	yeaL	fetR	fetL	falU	falD
INITIAL POSE: Pi	solU LEFT <--	76	71	66	66	68	75	70	68	67	76
	solU RIGHT -->	73	71	72	60	74	66	75	64	68	76
	solU INPLACE	K								55	
	solD LEFT <--	70	74	71	71	73	69	76	66	76	68
	solD RIGHT -->	70	73	65	62	65	63	65	71	74	65
	solD INPLACE		K								53
	logR LEFT <--	74	67	65	70	69	71	65	63	75	63
	logR RIGHT -->	60	60	67	70	69	70	64	74	59	76
	logR INPLACE			K		65		73			
	logL LEFT <--	55	66	70	63	67	66	68	65	65	80
	logL RIGHT -->	71	71	72	65	68	67	70	68	70	59
	logL INPLACE				K		64		71		
	yeaR LEFT <--	71	65	71	66	70	71	76	74	73	75
	yeaR RIGHT -->	57	69	69	63	70	67	73	72	64	72
	yeaR INPLACE			71		K		70			
	yeaL LEFT <--	52	73	73	71	70	71	69	71	59	58
	yeaL RIGHT -->	71	63	71	69	71	71	73	73	70	61
	yeaL INPLACE				70		K		71		
	fetR LEFT <--	67	62	68	73	62	60	75	61	68	67
	fetR RIGHT -->	58	73	69	73	67	62	o	g	b	r
	fetR INPLACE			74		68		K			
	fetL LEFT <--	67	71	71	69	68	73	65	72	66	63
	fetL RIGHT -->	71	58	76	67	70	65	G	72	72	65
	fetL INPLACE				70		71		K		
	falU LEFT <--	62	61	63	70	65	75	73	76	73	67
	falU RIGHT -->	65	63	71	66	70	69	76	59	74	65
	falU INPLACE	58								K	
	falD LEFT <--	73	74	78	77	73	65	71	62	79	79
	falD RIGHT -->	71	74	61	62	72	69	58	73	76	79
	falD INPLACE		54								K

(b)

Figure 4.14: MASH pose transition analysis in the DO mock-up ICU.
The analysis reports on detection and classification mean accuracy in the mock-up ICU under DO scene for singleview (a) and multiview (b) configurations.

trace indicates the predicted labels. Table 4.5.3 shows the pose descriptions used in the summarization plots. MASH's summarization results for fast motion of four patients are shown in Figure 4.16(a) using a singleview and (b) using a multiview configuration.

Comparison with Popular Methods.

The performance of MASH is compared with C3D [78] and the summarization and detection performance is shown in Figures 4.11 and 4.12. The sequence-overlaps achieved by each method in the mock-up ICU and the medical ICU are: 46.4% and 35.8% for conventional HMM, 70.5% and 63.3% for C3D, and 83.2% and 80.1% for MASH, respectively.

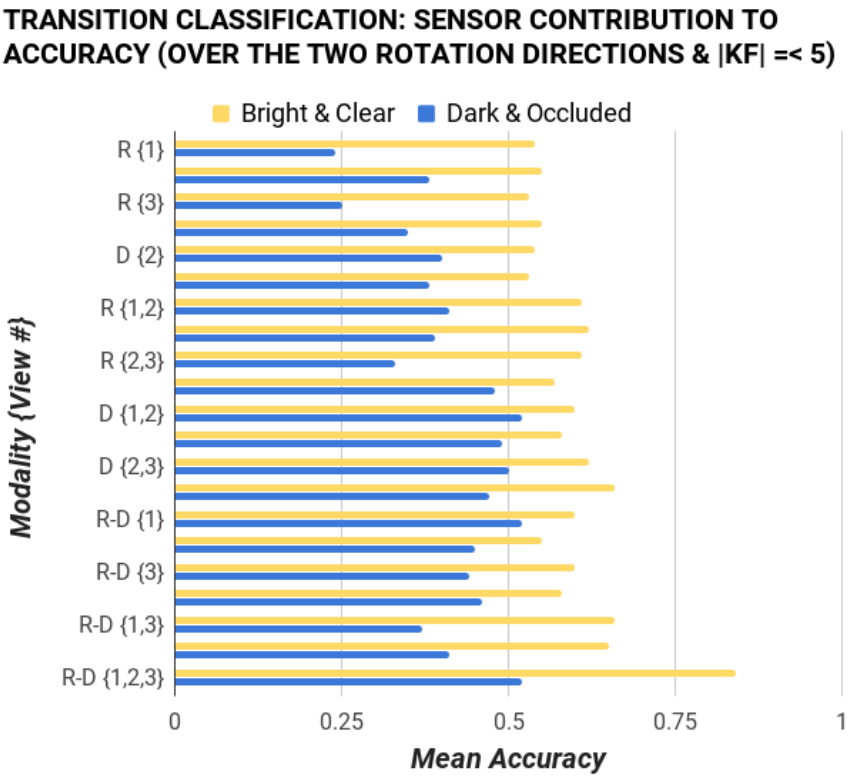


Figure 4.15: Sensor contribution to mean classification of transitions.
The MASH sensors and views are tested in BC (blue) and DO (yellow) scenes. The RGB and Depth modalities are represented by R and D, respectively.

Using a combined average, MASH outperforms HMM by 20% and C3D by 15% overlap.

Online and Offline Processing Speeds.

The **online** performance of MASH includes collecting data at 12 fps using the RPi3 devices. Real-ICU data collection is very critical and the main objective of the deployed system. Each device controls two modalities and synchronizes the data collection over the three views. RPi3s are incapable of extracting Inception features; therefore, feature extraction is defaulted to gMOM and HOG vectors as described in [75]. The performance is then extrapolated from the data collection to optic-flow computation and feature ex-

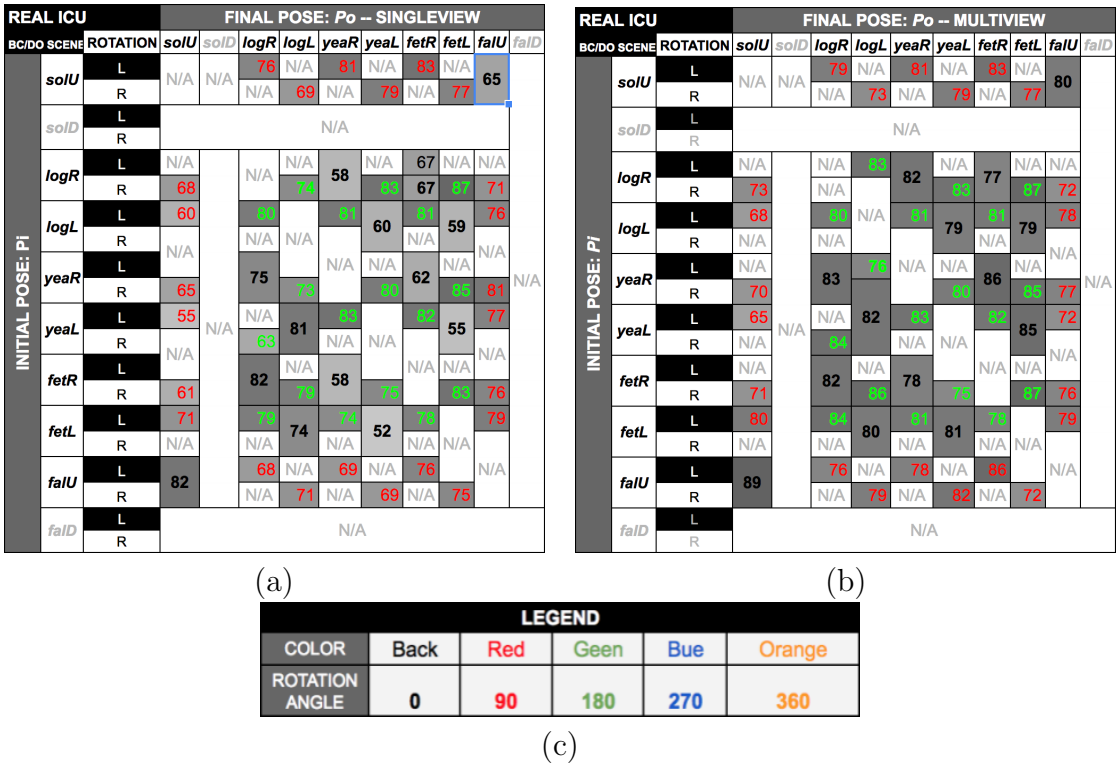


Figure 4.16: Pose transition analysis in the medical ICU.
The results report mean accuracy scores are shown for the singleview (a) and multiview (b) configurations with scale in shown in (c). The set of poses is reduced due to patient's constrained mobility and unavailable poses are marked N/A. Scene conditions are not quantized or controlled since the priority is patient care.

traction. The average running speed is approximately 6 fps: 12 fps for data collection (with pre-buffering), a drop of $2-3$ frames for optic-flow and a similar drop for feature extraction, with a under one frame for inference. The **offline** performance is extrapolated using desktop computers with GPUs to process the data frames and extract capable of extracting Inception features. The data collection in the mock-up ICU achieves 30 fps (with pre-buffering). Offline run-time performance is approximately 23 fps: 30fps video with a drop of 3 frames for optic-flow computation and 3 frames for Inception features and 2 fps for inference resulting in an average performance 22fps, which is just under four times faster than the simulated online approach using RPi3s. The summarization

results in Figure 4.12 use the offline approach.

4.6 Summary

Current computational abilities can help address the challenges of today's healthcare system. The proposed MASH framework enables the unobtrusive and non-disruptive data collection and analysis. The solutions ~~application~~ have the potential to improve patient care, develop new techniques, and objectively evaluate and validate medical treatments. The MASH framework is such an example. It can analyze patient poses in healthcare environments. Thorough evaluation highlights the feasibility of the detection and quantification of patient poses and motion dynamics for healthcare applications. The *mm* sensor network is robust to variations in illumination, view, orientation, and partial occlusions. MASH is non-obtrusive and non-intrusive, but not without a cost, as the patient-motion monitoring performance of MASH in dark and occluded scenes is far from perfect; however, most medical applications that analyze motion, such as physical therapy sessions, are carried out under less severe conditions. Although the deployed version of the system suffers slightly from under-powered devices, the findings reported in this chapter enable new studies and optimization opportunities.

Future studies will focus on system optimization. Also, future studies will investigate the analysis and recognition of activities and events in the ICU, such as hand sanitation. The continuous growth of the MASH dataset will enable deep learning analysis. An important future study will incorporate additional modalities, such as thermography, to validate findings and close the learning loop. Finally, effective medical applications require generating semantically meaningful logs. MASH will explore natural language understanding to create such logs and narrate ICU activities and events.

Chapter 5

Role Representation from Appearance and Interactions

The day science begins to study non-physical phenomena, it will make more progress in one decade than in all the previous centuries of its existence. -N. Tesla

5.1 Introduction

In this Chapter we expand the applications of the sensor network beyond pose analysis. Person identification, re-identification, and tracking are essential in many healthcare settings. However, collecting and using such identifiable information is prohibited in most cases by the Health Insurance Portability and Accountability Act (HIPAA) [16]. To address this limitation, this work introduces a novel framework for Role identification from Activity Maps (RAM). We demonstrate its application in an Intensive Care Unit

setting in a real hospital environment where RAM learns ICU-associated roles. RAM combines static appearance features (texture and color) with quantified dynamic human locations and interactions (semantic maps) to describe these roles. The problem of person identification and re-identification depends on two main aspects: identity representation and discriminant metric definition [39]. An effective identity representation is robust to changes in illumination and view point, and allows for effective matching in multiple instances. A reliable metric allows for clear distinctions and can be used for multi-target tracking, identification, and re-identification tasks. The overview of RAM is shown in Figure5.1, where the main components include the scene initialization, plane estimation, object detection and localization, person detection (compute appearance feature f_A) and interaction-histogram initialization, person poselet quantization and time-evolving interaction computation (T), for view ($v \in V$). The output is the interaction histogram, which is the semantic feature vector $f_{M,T,V}$.

The proposed representation is simple, easy to compute, and robust to natural conditions, which makes it suitable for low-power distributed network deployment. The semantic maps make RAM independent of network configuration. The performance of RAM is evaluated on 11 days of multimodal multiview data and compared with the latest methods. Thorough evaluation of RAM is performed to justify its components and to compare its performance with competing appearance-based and tracking-based methods. The findings from this approach will enable the privacy-compliant analysis of workflows in healthcare and other areas where identifying individuals is not permitted. RAM identifies roles (not individuals), protecting patient and staff privacy, while ensuring workflows remain unaffected by surveillance mechanisms. Figure 5.2 shows sample inputs, detected relevant objects, and estimated roles in a mock-up and a medical ICU room.

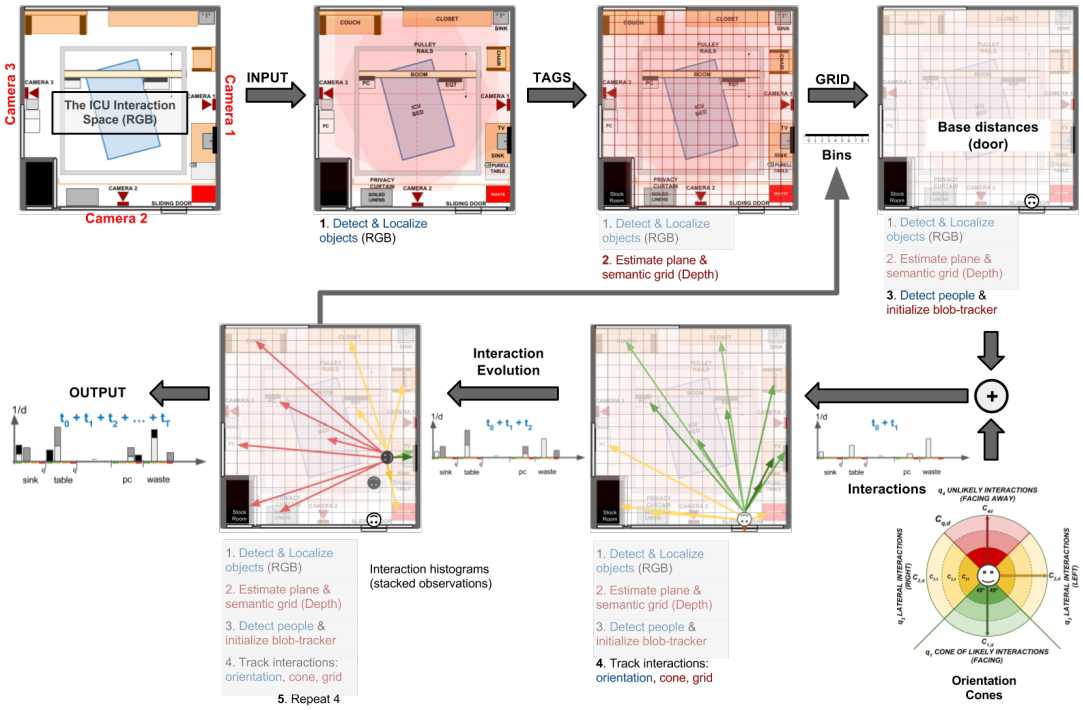


Figure 5.1: RAM overview: role representation and feature computation steps. The overview is depicted in clock-wise order. The input scene is selected and the cameras are positioned to observe the space. Then the objects are detected using RGB data. The floor plane is estimated and a grid is computed and overlayed to localize objects using the depth and distance maps. Then when a person is detected (f_A is computed) the base distances (i.e., DC-offset) are computed relative to the person being detected on the entry way and used to initialize a histogram. The interaction cone is then used to quantize poselet-orientations and being populating the histogram as the visit interaction evolves over time (T) across all views (V). The output is the interaction histogram or feature $f_{M,T,V}$.

5.1.1 Medical Background

There is an increasing interest in role identification and analysis in healthcare [73], due to its potential benefits in improving and optimizing care. One major gain from role identification and analysis is in defining each person's responsibilities, ensuring appropriate implementation of each professional's role, optimizing professional scopes of practice, and thereby ensuring efficient patient management [7]. Although clinicians agree that

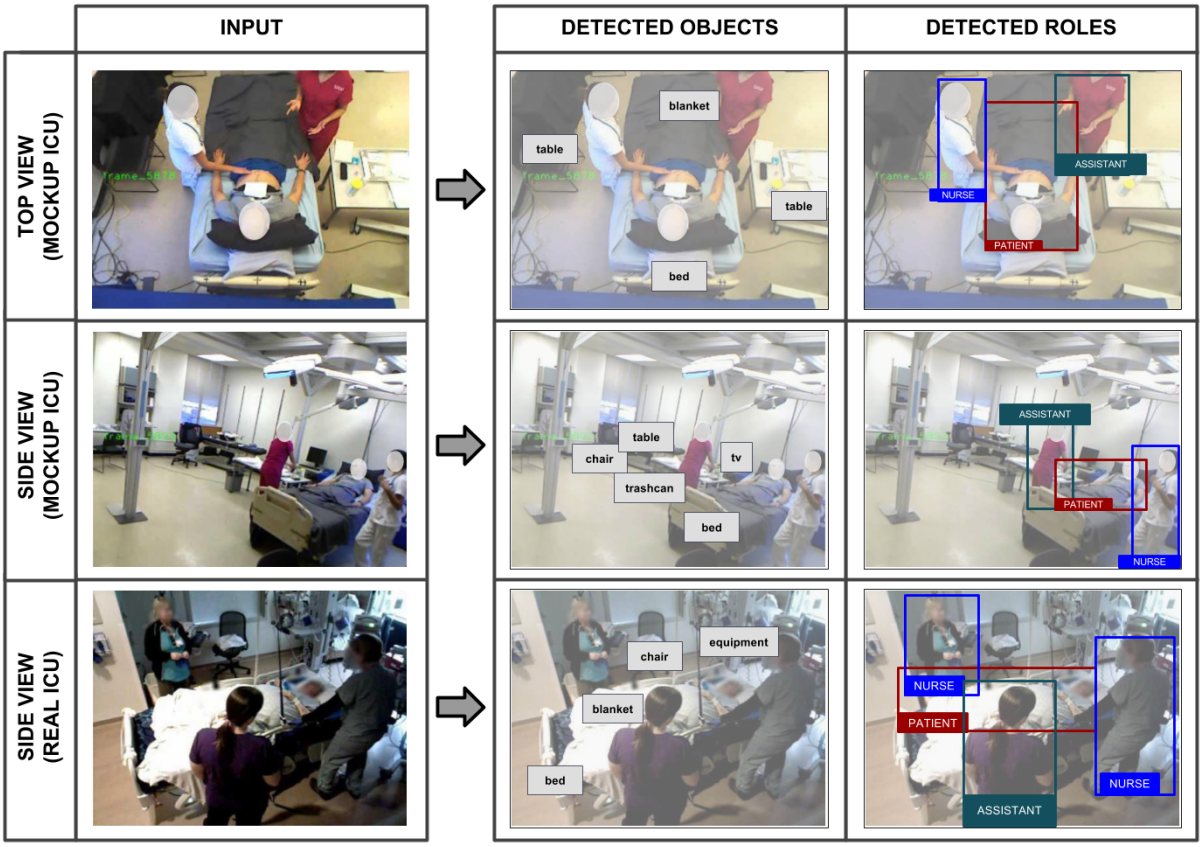


Figure 5.2: RAM sample detected roles in the ICU rooms.
The mock-up ICU room is shown on top two rows and the medical ICU room is shown on the bottom row. The columns are the input frames (left), the detected objects (center), and the detected roles in the color bounding boxes (right).

detailed understanding of workflows is essential to quality of care, healthcare restrictions prohibit the use of people's identifiable information. To circumvent these data restrictions, RAM introduces methods for role representation and role identification based on appearance features (for training and system initialization) and semantic activity maps (location and interactions).The idea behind semantic maps is that different roles interact with different objects, visit different locations, and maintain a certain distance from certain objects in the ICU. This is further enhanced by observing subjects over time. For example, a patient might walk past a ventilator, but a medical practitioner will spend

more time in that area. Similarly, visitors spend more time close to the patient sitting in chairs, while staff mostly avoids chairs. Although semantic map features alone can be used to infer roles, their combination with appearance features achieves greater accuracy.

5.1.2 Technical Background

Studies analyzing healthcare environments include using a single RGB-D sensor, RFIDs, and proximity sensors to record activities in a neo-natal ICU as in [37] Workflows in an operating room are analyzed in [55] and the analysis tasks are very complex. One significant limitation is considering that any activity can be performed by any individual. This makes the action space relatively large, which decreases accuracy. One helpful concept in improving outcomes includes identifying roles who perform distinctive and common activities and using this information to identify roles (e.g., patient, doctor, or staff). The surveys from [80] and [32] describe the challenges and most popular techniques in person re-identification. Existing methods for identification and re-identification range from methods leveraging deformable parts [8] to feature representation and metric learning [39] to video ranking [82] (as an alternative to single-frame approaches). The work in [48] introduced a distributed network framework for node performance comparison and person re-identification that can be used to estimate optimal camera topology. The authors in [18] argue that most existing methods depend on person pose and orientation variations and introduce a technique to model such variations in the feature space. Also, there are several feature representations that have pushed the limits of performance to new levels. Appearance-based representations such as the ensemble of local features (ELF) [21] and symmetry-driven accumulation of local features (SDALF) [4] encode color properties. Similarly, salience matching and learning [89], [90] and mid-level filters [91] depend on relative patch contrast and distinctiveness. Although the previously cited

research achieves impressive results, their appearance-based methods directly depend on proper imaging conditions, such as bright, and uniform illumination, and view angle between the individual and the camera. In addition, appearance-based role representation alone is not sufficient. For instance, medical isolation procedures to protect compromised patients require that all people entering the ICU room wear disposable isolation scrubs, so all roles appear identical. Another limitation of these representations in real-world applications, such as healthcare, is their inability to evolve over time (i.e., to consider temporal information) and to integrate interaction information. The proposed approach introduces a novel role representation; a semantic activity abstraction and extraction algorithm to identify; and a method for role identification based on the sequence of observed activities, visited locations, and detected interactions (cones for orientation and proximity). The proposed methods are capable of dealing with cases when role-based visual features are obfuscated by extreme scene and appearance changes.

5.2 RAM Dataset

The sensor locations and camera views are shown in Figure 5.3. The methods from [24] are used to calibrate the three cameras and estimate the floor plane of the ICU. A total of eleven days of video data (approximately a total of 264 hours, 15,840 minutes, or 950,400 seconds) are collected covering six nurse assistants, four caterers, five medical doctors, four facilities and janitorial personnel, ten nurses, five patients, twelve visitors, and two days of isolation. Additional multimodal representations and views of the medical ICU room are shown in Figures 5.4 and 5.5.

The set of roles $\mathcal{R}$ and symbols representing each of the eight observed roles are: nurse (A)ssitant, (C)aterer, medical (D)octor, (F)acilities, (I)solation, (N)urse, (P)atient, (V)isitor. The role set is indexed by r, where R_r with $r = 1$ is used to indicate the role of

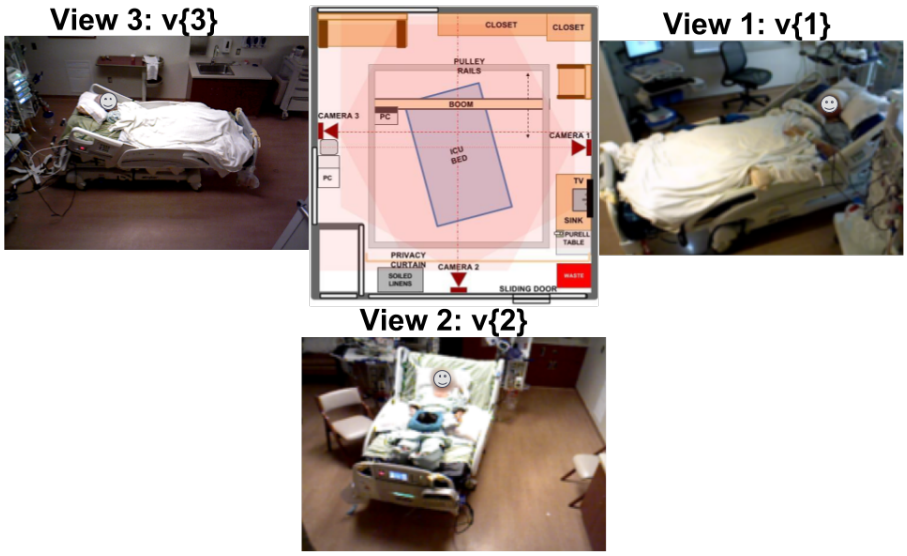


Figure 5.3: RGB-D camera locations and views of the medical ICU room.
The three nodes in the medical ICU room $(1, 2, 3)$. Each node is composed of one RGB-D sensor, one Raspberry Pi3, one $24,000$ mAh battery, and an aluminum enclosure, which contains all node elements.

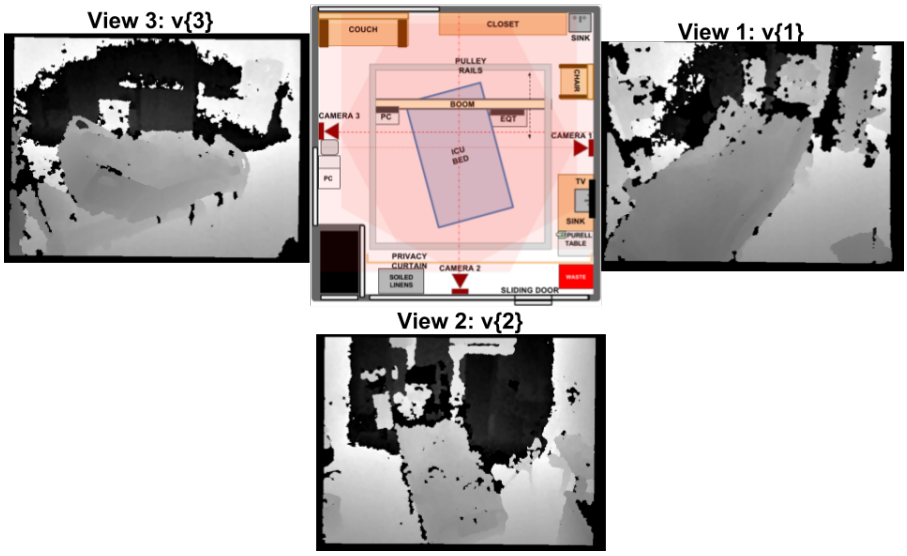


Figure 5.4: RAM depth views of the medical ICU room.
The three views of the medical ICU room $(1, 2, 3)$ shown the ranged and normalized imaged distance $[0, 255]$, where pixel intensity represents distances of points in the scene to the camera sensor.

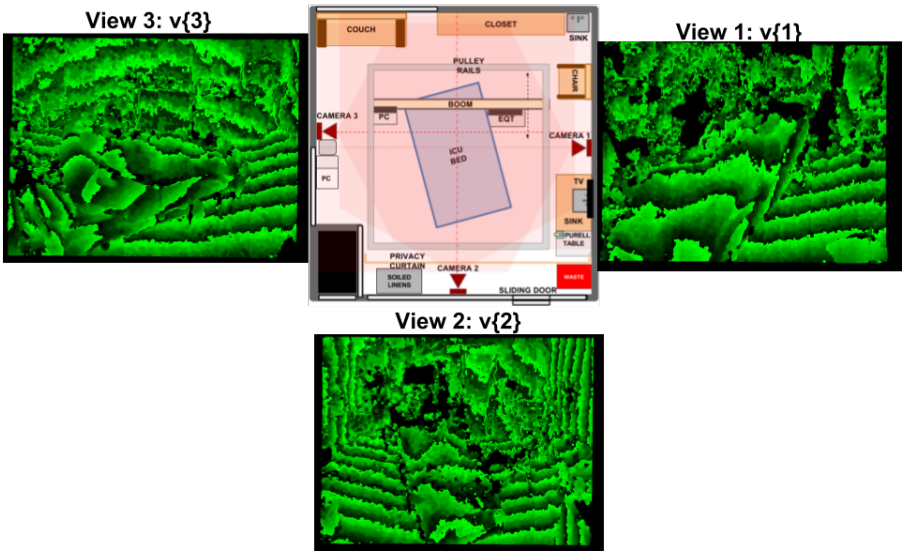


Figure 5.5: RAM distance views of the medical ICU room.
The three views of the medical ICU room $(1, 2, 3)$ shown the quantized distance as a three channel color image $[0, 2^{11}]$, which represents distances of points in the scene to the camera sensor.

nurse assistant. Figure 5.6 shows samples of the various roles. The total frequency count of observed/collected instances for each role (in number of minutes on the vertical axis) is show in Figure 5.7. The data also include three hours collected in a mock-up ICU room with actors playing four patients, one nurse, one visitor, and one nurse assistant. Note that about 30% of the data contains more than one role, patient-visitor being the most common. The scope of this work is focused on role representation and identification.

Figure 5.6: Eight roles associated with the ICU room.

Recall that when the room is in isolation, all visitors as well as hospital staff are

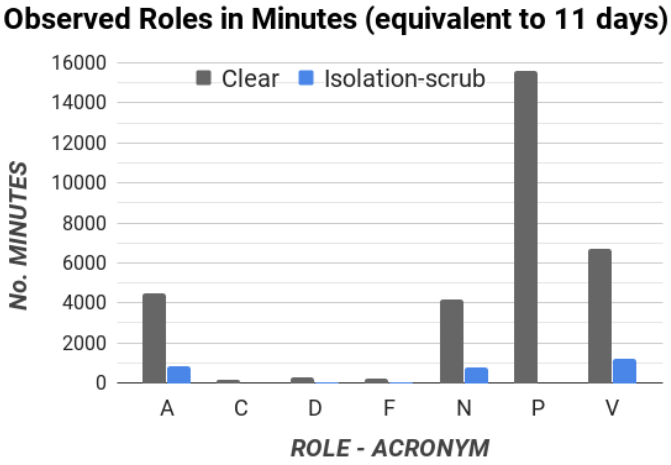


Figure 5.7: Number of minutes each of the roles is observed.
The data covers 11 days. Gray bars indicate that differentiating roles based on appearance is possible, while blue bars indicate isolation scrubs are used (i.e., no appearance difference). Isolation scrubs were observed for a total of 2,795.04 minutes or 1.91 days.

required to wear disposable scrubs, causing all roles to appear identical. Multiple roles were observed manually and added to the counts for these two days.

5.3 Description of the Problem

There are multiple problems and stages in role representation and role identification. At the high level, the objective can be described as assigning roles to people observed in an ICU room using multimodal and multiview videos. The challenges come at the lower levels of the analysis. For instance, what are good appearance role representations, how do roles evolve over time, and how can one infer roles when appearance features are not discernible. Therefore, given a set of training multimodal multiview videos, the first problem involves identifying a role using appearance features to create an appearance dictionary. Nurses tend to wear unique uniforms in order to keep a sense of their individuality in the work place [7]. This makes identification based on appearance alone

unreliable. The solution is related to the second problem, which corresponds to learning the semantic motion dynamics associated with each of the observed roles and creating a dictionary of such representations and roles. The last problem involves matching an unseen video with information extracted at training to find the best matching role R^* from matching appearance features to obtain the appearance and interaction scores for all roles in the set $\mathcal{R}$ = [Assistant, Caterer, Doctor, Facilities, Isolation, Nurse, Patient, and Visitor], indexed by $r, 1 \leq r \leq R$, from all views $v, 1 \leq v \leq V$, and across all frames $n, 0 \leq n \leq N$..

5.4 Approach

Consider the ICU rooms in Figure 5.2. Intuitively, hospital visitors look different from healthcare staff, often dressing differently. In addition, different roles perform different activities and visit different locations in the room. Some activities such as entering the ICU are performed by all roles, while some social or medical activities are performed only by specific roles. For example, nurses check ventilators and janitorial personnel clean rooms and empty trashcans, while visitors sit and interact with patients for longer time intervals and in closer proximity. The objective is to identify the set of locations (corresponding to the various activities) that each role visits along with the associated objects and interaction cone configurations. The interaction cones are used to quantify a person's relative distance and orientation to objects of interest. The variability of visit duration by a role and the set of observations (semantic features) is unbounded and can be very short or very large. A method to extract semantic features over time to deal with the variability in visit duration by certain roles is also proposed. The features are discriminative and informative and their nature allows them to be independent of their chronological order.

The first step is detecting individuals in the scene via [10]. The set of interaction features representing the r-th role ($R_r \in \mathcal{R}$) at frame n is $\zeta_{n,r,v} = \{C_{o,q}\}_{O,Q}, 1 \leq o \leq O1 \leq q \leq 4$, where single element $C_{o,q} \in \{1, 2, 3\}$ is the interaction cone vector for the o-th object at with relative orientation q and interaction distance d. The set of tagged objects is $\mathcal{O} = \{$bed, chair, computer, doorway, person, sink, table, trashcan, tray, and ventilator$\}$ indexed by o for $1 \leq o \leq 10$. Objects are detected using the RGB modality and localized using the Depth modality.

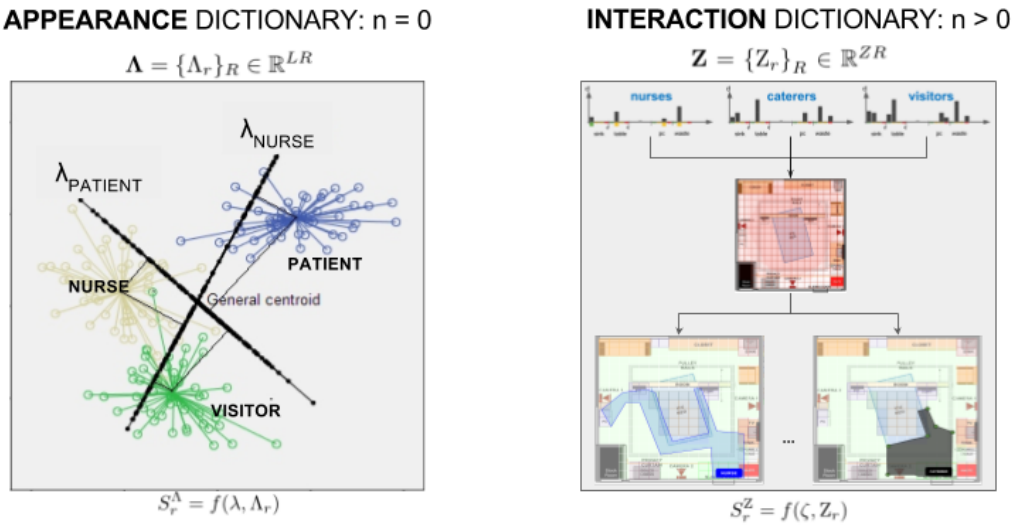
5.4.1 Training

Training is done in two stages: static, where appearance features $\lambda_{n,r,v}$ computed at $n = 0$, serve to create the dictionary of role appearances $\mathbf{\Lambda} = \{\Lambda_r\}_R$; and dynamic, where a dictionary of role-object interactions $\mathbf{Z} = \{\mathrm{Z}_r\}_R$ is constructed using the interaction features $\zeta_{n,r,v}$ for all roles and across all frames and views.

Role Representation

We use role representation from appearance and interaction information to deal with these ICU restrictions. It assigns roles over the complete activity or event using a threshold (70%) based on the number of frames or observations to link a role, else the role is considered to be "unknown". The process of learning a role representation starts with identifying the set of interactions corresponding to each role. Given a set of videos $\mathcal{F} = \{F_k\}_K$, each with N-frames, $0 \leq n \leq N$, the first step is to extract the appearance. The appearance vectors ($\lambda_{n,r,v}$) are computed at $n = 0$ and used to construct the dictionary of appearances for all roles $\mathbf{\Lambda} = \{\Lambda_r\}_R$. Similarly, the interaction vectors ($\zeta_{n,r,v}$) are computed for $1 \leq n \leq N$ and are used to construct a dictionary of role-interactions for all roles $\mathbf{Z} = \{\mathrm{Z}\}_R$. Figure 5.8 shows the two dictionaries: (a) appearances and (b)

interactions.



(a) (b)

Figure 5.8: RAM appearance (a) and interaction (b) dictionaries.

Appearance Dictionary ($\mathbf{\Lambda}$)

Appearance vectors $\lambda_{n,r,v}$ are computed for each person as they enter the ICU room (i.e., frame $n = 0$) using the data from available view v. The vectors computed have two parts: a 128-dimension GIST vector (one scale) for texture [54], and the 96-dimension (first and second order) color histogram vector [88] by combining the first moment (mean) and second moment (standard deviation) on 16-bin histograms extracted from each of the three channels in the HSV color space. The texture and color features are concatenated to form the vector λ. The intuition is that these vectors can help identify distinct visitor clothing patterns and generic healthcare staff uniforms. These vectors are used to create the appearance dictionary $\mathbf{\Lambda} = \{\Lambda_r\}_R \in \mathbb{R}^{LR}$, where $L = |\lambda| = 224$ is the cardinality of the appearance feature vector and $R = 8$ is the number of roles. The elements of the dictionary $\mathbf{\Lambda}$ are the Linear Discriminant Analysis (LDA) [57] boundaries for each role,

each represented by Λ_r. The decision hyper-planes are used to score a new sample by computing the distance to all, but selecting the closest one.

Location and Interaction Quantification The semantic interaction map is combined the interaction cone to localize individuals and objects and compute the semantic interaction vectors. A sample, empty semantic interaction map is shown in Figure 5.9. A set of maps is computed for each of the eight roles over time. The results in section 5.5 show two computed interaction maps.

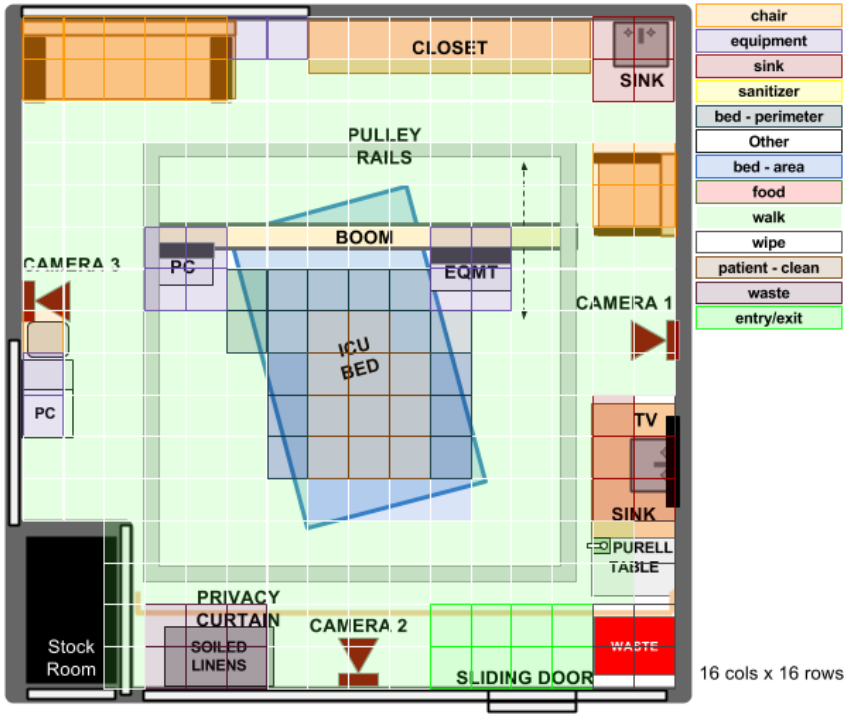


Figure 5.9: RAM semantic interaction map.
The map is constructed for role computation in a 16x16 grid overlayed in black.

The interaction cone is shown in Figure 5.10; it represents the 40-element vector $C_{o,q}$ for $1 \leq q \leq 3$ and $1 \leq d \leq 4$, with shape $[c_{o=1,q=1}, \ldots c_{o=1,q=4}, \ldots c_{o=10,q=4}]$. The feature vector is computed for each detected individual at frame n. The values are the distances, and the vector element indexed by the object and orientation quadrant.

The distances values are assigned based on average arm length, where $1 : close(< 2ft)$, $2 : nearby(> 2, < 4ft)$ and $3 : far(> 4ft)$. The poselets from [6], [47], and [62] are evaluated for usage in the ICU. Experimentally, the poselet estimator from [62] is used to compute the orientation cones, which is assigned to the closest quadrant. The poselets are given with respect to the camera that detected the person and mapped across the ICU floor plane. The distance (disk) is computed between detected blobs (objects and individuals) centroids and assigned to the closest d.

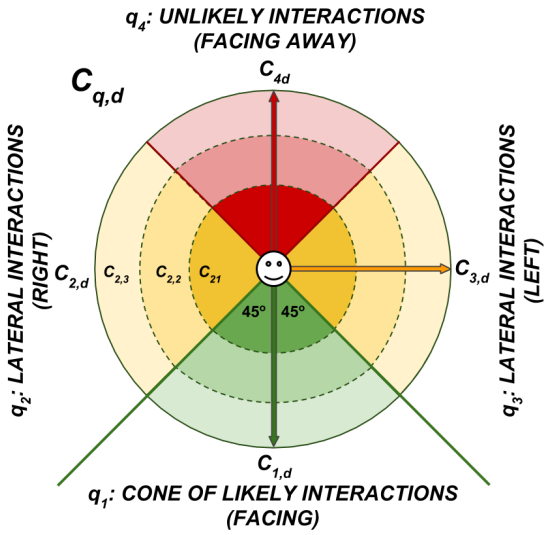


Figure 5.10: RAM role interaction quantification cones.
The cone elements are $C_{q,d}$ for quadrant q and distance d. The elements indicate regions of highest interaction (green), mid interaction (yellow), and lowest interaction (red). Darker disks indicate higher interaction probability (closer), while lighter disks indicate lower interaction (farther away). The color and color scales are used to indicate person orientation and proximity.

Interaction Dictionary (Z)

The interaction features representing the r-th role at frame n correspond to the interaction cones (i.e., $\zeta_{n,r,v} = \{C_{o,q}\}_{n,r,v}$) computed for each role r and each available view v at frame $n > 0$ from a network with V views over a total of N frames. The floor plane is

estimated from the depth modality to localize 10 tagged objected and compute interaction features, which are person-object relative distances and orientations. The interaction vector is noted as x_ζ and has 40-elements representing four relative orientations to each of the 10 objects. This vector represents the evolution of roles interacting with ICU objects over time. Interaction features vectors are clustered using density based clustering (DBSCAN) [12] and the resulting in the interaction dictionary $\mathbf{Z} = \{\mathrm{Z}_r\}_R \in \mathbb{R}^{ZR}$, where each $\{\mathrm{Z}\}_r$ represents the cluster centroid for role r, $Z = |\zeta| = 40$ is the cardinality of the interactions feature, and $R = 8$ is the number of roles.

The set of images in Figure 5.11, shows the evolution of the semantic interaction map features as individuals appear and "interact" with ICU objects in the scene. The figure shows the interactions of an unknown role. The main steps are: scene initialization (RGB object detection); floor plane estimation, object localization, and grid overlaying (Depth and Distance maps); detect and track people (RGB and Depth blobs); track person orientations and locations, compute relative person-object distances over time (i.e., build the interaction vector).

5.4.2 Testing

The objective at testing is to find the role R with the maximum score across all views and frames ($0 \leq n \leq N$). The average score S_r^Λ for role r is computed for a new individual at $n = 0$ using available view v via:

$$S_r^\Lambda = \frac{1}{V} \sum_{v=1}^{V} \mathrm{D}(\lambda_{r,n,v}, \Lambda_r), \forall v, \forall r, \tag{5.1}$$

where $\mathrm{D}(\cdot)$ is the Euclidean distance computed between the input appearance vector $\lambda_{n,r,v}$ and each role boundary $\{\Lambda_r\}_R$.

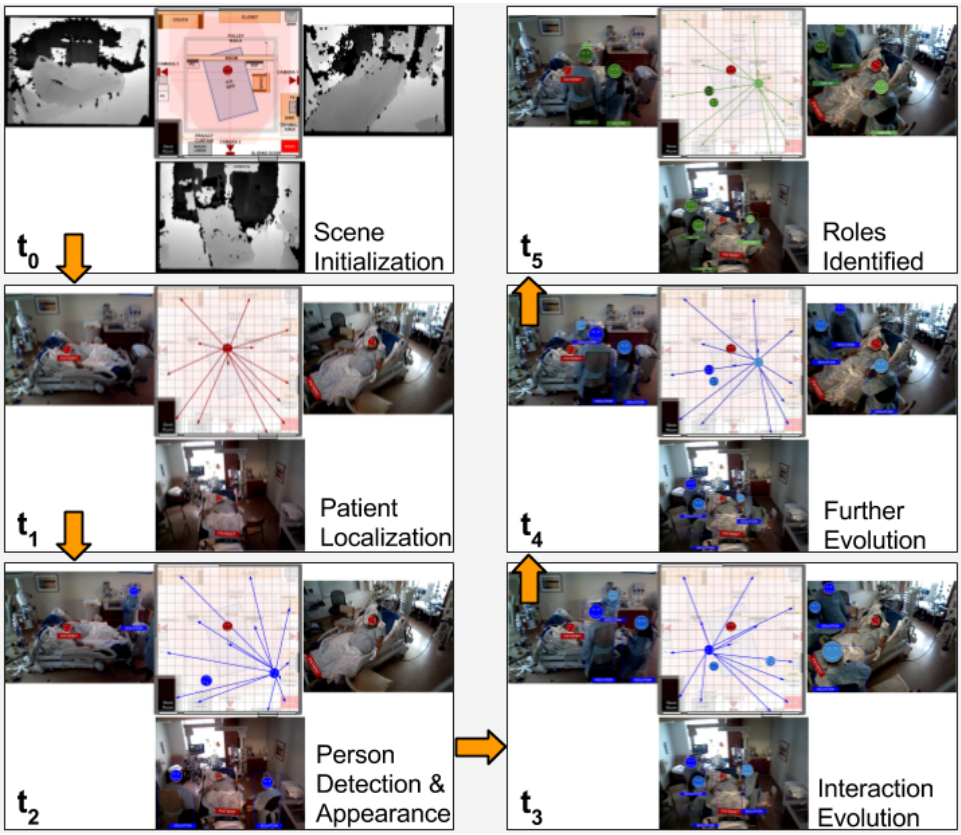


Figure 5.11: Semantic feature evolution for an isolated ICU room based on interaction maps. From left-to-right and top-to-bottom roles res are estimated from videos, in this example, in six time-samples. At t_0 the scene is initialized and objects are localized and tagged using multimodal multiview information. At t_1 the patient is localized within the room. At time t_2 the appearance and interaction features of individuals entering the room is computed. The role features evolve from t_3 to t_4 based on the individual-to-object distances and orientations Finally, at time t_5 the most similar roles are selected based feature histograms for each person in the video.

The interaction scores are computed for $n > 0$ via:

$$S_r^{\mathrm{Z}} = \frac{1}{V} \sum_{v=1}^{V} \sum_{n=1}^{N} D(\zeta_{n,r,v}, \mathrm{Z}_r), \forall r, \tag{5.2}$$

where $\mathrm{D}(\cdot)$ is the distance between the interaction vector $\zeta_{n,r,v}$ and the role-centroid $\mathrm{Z}_r \in \mathbf{Z}$ at frame n from view v.

Appearance and Interactions for Role Identification

Role candidates $S_r, 1 \leq r \leq R$ are a combination of an individual's appearance and interaction scores:

$$S_r = (S_r^{\Lambda} + S_r^{\mathrm{Z}}) \tag{5.3}$$

The estimated role R^* is the one with the most similar representation over all roles given by:

$$R^* = \arg \min_{1 \leq r \leq R} (S_r) \tag{5.4}$$

This approach has the additional advantage of ignoring the sequence of activities, which are not required to be sequential.

5.5 RAM Experimental Results

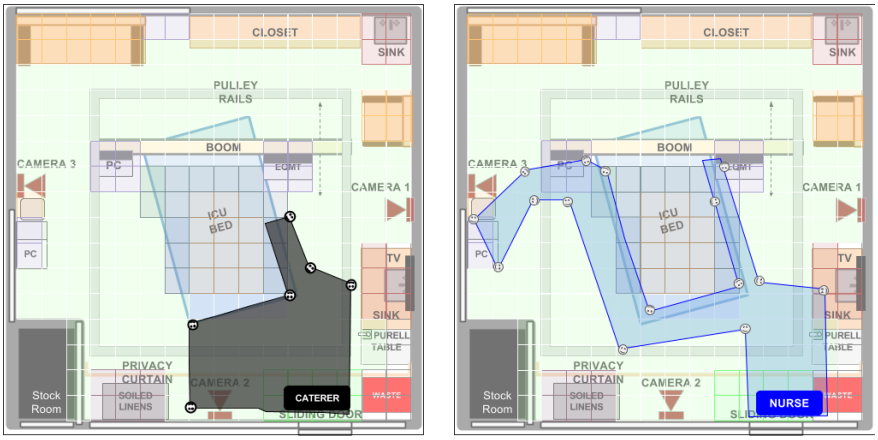
The performance of RAM is evaluated under different camera views for accurate role identification using a stratified 10-fold cross-validation evaluation scheme. Average results across all folds are presented.

5.5.1 Semantic Interaction Maps

The maps in Figure 5.12 shows interactive maps constructed for the caterer (a) and the nurse (b) roles.

5.5.2 Non-Isolated and Isolated Environments

This experiment contains two parts. The first part uses appearance features and semantic maps for role identification in non-isolated environments. The confusion matrix



(a) (b)

Figure 5.12: RAM semantic interaction maps.
The two maps are constructed for (a) the caterer and (b) the nurse roles overlayed in black and blue, respectively.

in Figure 5.13 (a) shows the qualitative performance of the proposed role representation. The second part of the experiments takes place in an isolated environment, where individuals have to wear blue disposable scrubs and appearance features are informative but non-discriminative. This means that appearance is used to detect the blue scrub but not to identify roles. In such case, the semantic features become the relevant input for classification and identification of roles. The confusion matrix for isolated scenarios is shown in Figure 5.13 (b).

The role classification accuracy as a function of the number of semantic features observed over time ($t > 0$) is shown in Figure 5.14. The traces represent a one-vs-rest scheme for each role. The contribution of appearance features (λ), semantic features (ζ), and their combination for role identification is shown in Figure 5.15.

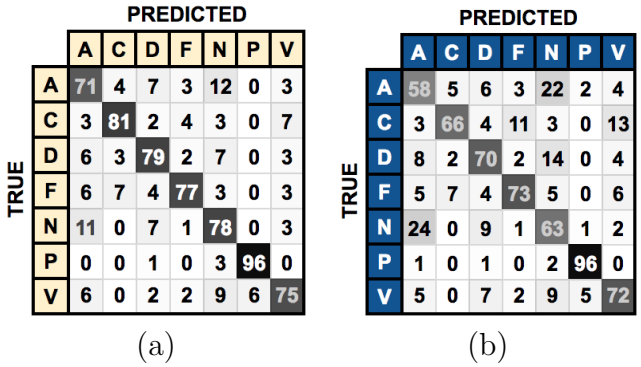


Figure 5.13: Role identification confusion matrices.
(a) confusion matrix of non-occluded identified roles. (b) Confusion matrix of the identified roles for individuals wearing isolation scrubs. The role symbols are A: assistant, C: caterer, D: doctor, F: facilities, N: nurse, P: patient, and V: visitor.

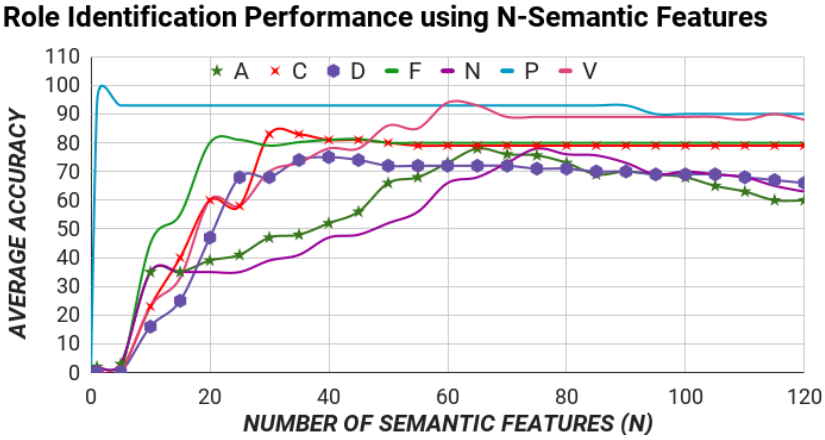


Figure 5.14: Mean role classification dependency on number of semantic features.
The performance depends the number of extracted semantic features (i.e., over time). The duration of the observations helps to identify roles. Overall, roles observed for shorter periods of time are harder to identify. The vertical dashed line indicates that 74 observations is on average the best number for all detections.

5.5.3 Decentralized Process: Camera Views

This experiment evaluates individual views and combinations of views by modifying equations 5.1 and 5.2. It serves to identify optimal views for accurate role identification. Obtaining a clear (unobstructed and direct) view of the activities and roles directly

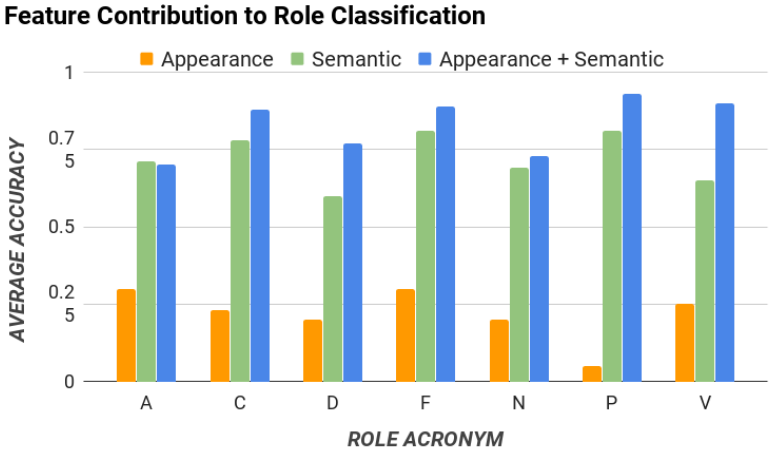


Figure 5.15: Feature contribution to mean role classification accuracy.
using appearance features (f_A) as orange bar, semantic features (f_M) green bar, and their combination ($f_A + f_M$) as blue bar. The role symbols are A: assistant, C: caterer, D: doctor, F: facilities, N: nurse, P: patient, and V: visitor.

affects the role identification results shown in Figure 5.16. Camera locations and views are shown in Figure 5.3. There are two objectives behind this experiment: the first is to show that the decisions can be made at the individual nodes; and the second is to explore the best and worse case scenarios and to simulate the effects on identification performance due to sensor failures or sensor occlusions (i.e., in the ICU, views can be blocked by privacy curtains).

5.5.4 Multiple-Target Role Identification

This experiment uses the combined RAM elements to represent, track, and identify roles in the ICU. The experiment is performed on video instances with one or more people present in the scene. This experiment justifies the dimensionality of the map using accuracy and complexity as shown in Figure 5.17.

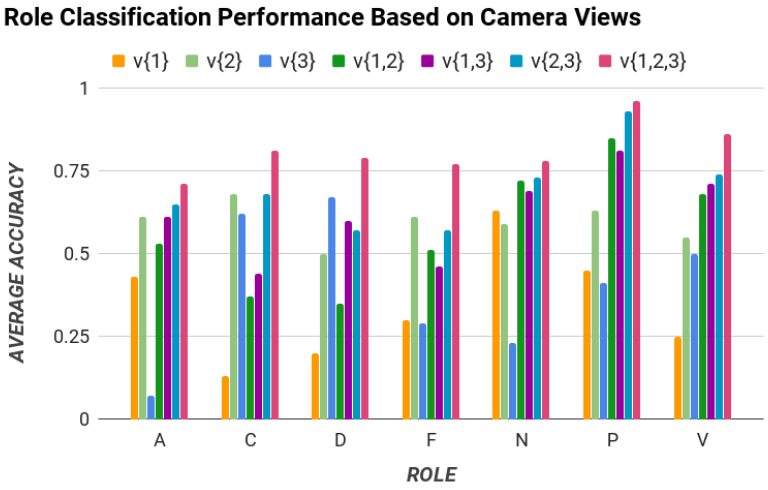


Figure 5.16: Average classification performance per view(s).

using all views and various reduced camera-view combinations. Sensor locations and views from right to left in clock-wise direction are: v{1}, v{2}, and v{3} shown in Figure 5.3. The bar plots indicate that better views of locations visited by specific roles help to better identify roles, while the best performing combination is the complete set of views. The role symbols are A: assistant, C: caterer, D: doctor, F: facilities, N: nurse, P: patient, and V: visitor.

5.5.5 Performance Comparison

The performance of RAM is compared with competing state-of-the-art methods as shown in Figure 5.18. The contrast methods are You-Only-Look-Once (YOLO) [63] and the method based on deformable part models and appearance from [39]. However, the competing methods only apply to the non-isolated environments, where appearance can be used to identify roles. The methods can detect isolation scrubs but cannot identify the occluded role.

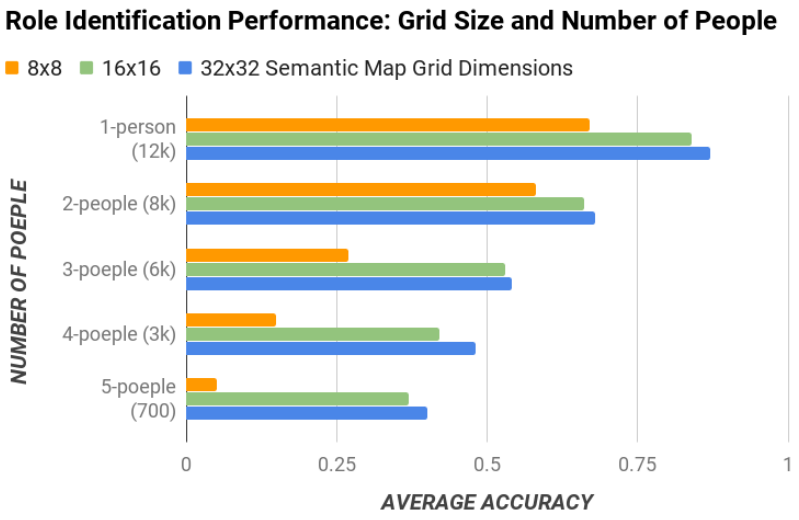


Figure 5.17: Effects of grid size on average role classification.
The accuracy as function of semantic map grid dimensions and number of people in the scene. The worst performing grid size is 8×8 due to the artifact in which multiple people can occupy the same grid. The best performing grid size is 32×32; however, it is also the more complex.

5.6 Summary

The classification results indicate that roles can be identified by using a combined appearance and semantic activity approach. In some cases, individuals can be identified at the moment of first detection ($t = 0$) based on appearance features only. Although decentralized decisions are possible at the node level, the best individual decisions depend on having the optimal view of the activities in the room. The best role identification performance is achieved when appearance and semantic data from all nodes are combined.

The grid dimensions are evaluated experimentally. The best compromise between complexity and performance is met with the 16×16 grid size, as shown in Figure 5.17. In this case, each grid covers an approximate area of one square foot, which coincidentally is also the average area covered by a standing person (scene's top-view). Recall that the objective of RAM is to identify various specific roles in the ICU. To reduce the study's

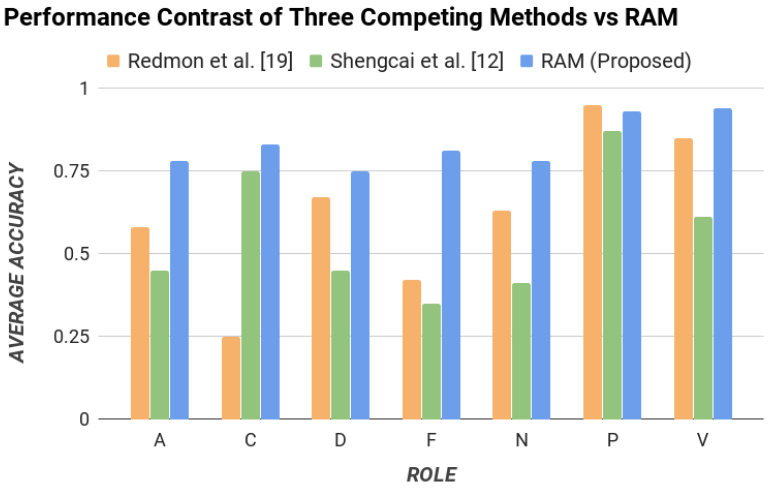


Figure 5.18: Performance of RAM compared with two methods.
Four competing methods and RAM for ICU role identification in clear and isolation (I) conditions. The role symbols are A: assistant, C: caterer, D: doctor, F: facilities, N: nurse, P: patient, and V: visitor.

search space, re-ranking was explored, but it was omitted due to its minimal impact. The intra-class similarities (e.g., among nurses, nurse assistants, and doctors) are small, compared with the large inter-class differences (e.g., between medical and non-medical staff).

Possible future directions explore the evolution of roles in healthcare. For instance, due to the scarcity in the healthcare workforce, regular ICU visitors are often trained on basic care tasks, alleviating some of the task load on the staff. This evolution can obfuscate RAM's re-identification analysis and can have a negative impact in its performance. Other possible directions will investigate the identification of new roles based on anomaly detection.

Experiments involving object detectors and human poselet estimators indicate that good object detectors and poselet estimators directly affect the performance of RAM. Although multimodal-based detectors are still in their infancy, they are shown to outper-

form unimodal (e.g., RGB-only) object detectors. One improvement in the system and and future area of research would be the incorporation of object detectors and poselet estimators that combine RGB-D data. RAM uses a predefined set of objects; however, these directly depend on the specific environment and application. Future studies will explore identifying relevant objects and estimating an object's importance in role representation and identification. It is important to note that not all roles are explored in this study due to the limited number of observed instances. The continuous expansion of this study will allow the integration and analysis of additional roles. High-level semantic activities are used in this study. However, a possible future directions should focus on finer analysis to better infer roles and simultaneously tackle activity recognition.

Chapter 6

Healthcare Activity and Event Analysis

The present is theirs; the future, for which I really worked, is mine. -N. Tesla

6.1 Introduction

This chapter tackles the analysis of activities and events in the ICU. It introduces the Healthcare Event Analysis and Logging (HEAL) framework, which focuses on the detection of human activities and their sequential ordering to create event logs and classify events into four. The main novelty of HEAL is the creation of chronologically consistent event logs by fusing contextual and visual information from multiple views and modalities. Contextual information includes location, relevant scene objects, duration of activities or events. We introduce the concept of actor roles, i.e., individuals present in the scene

are identified based on their interactions as opposed to recognizing their identities. This is especially important given the ICU conditions and generally accepted protocols for security and privacy of patients and staff in such environments. Figure 6.1 shows the overall event analysis workflow consisting of three stages: aspect initialization, aspect computation, and label estimation.

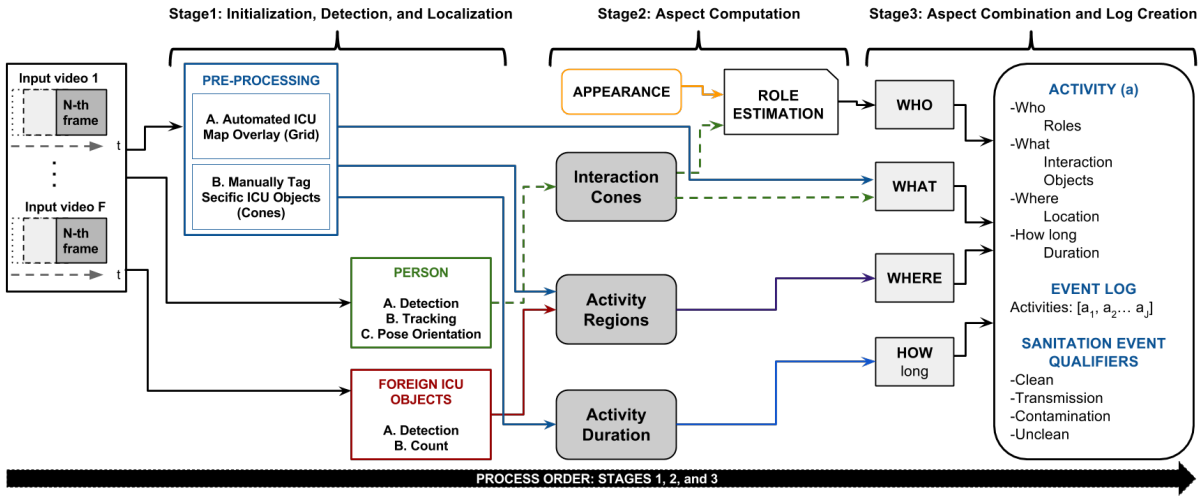


Figure 6.1: Contextual aspects stages for activity and event analysis.
Stage 1: estimation and overlay of the activity map on the ICU space, tagging and localization of ICU-objects; detection, tracking, localization of people and objects (foreign to the ICU) on the map. Stage 2: computation of interaction cones as individuals and ICU-objects relative orientations and distances, identification of the activity grid, and estimation of activities duration. Stage 3: combination of aspects to create logs, estimation of activity labels, localization of activities in time, and computation sanitation-event qualifiers.

We use the MESH network from Chapter 4 to collect the data used in the analysis and logging of human activities and events. The multimodal sensor nodes are installed at various locations inside the ICU room to monitor the space from multiple views, see Figure 6.2 for a top-view in an ICU space. The multimodal multiview nature of HEAL allows it to accurately monitor the ICU room and is robust to scene conditions such as illumination variations and partial occlusions. HEAL is currently deployed in a medical

ICU where it continuously monitors two rooms without disrupting existing infrastructure or standards of care.

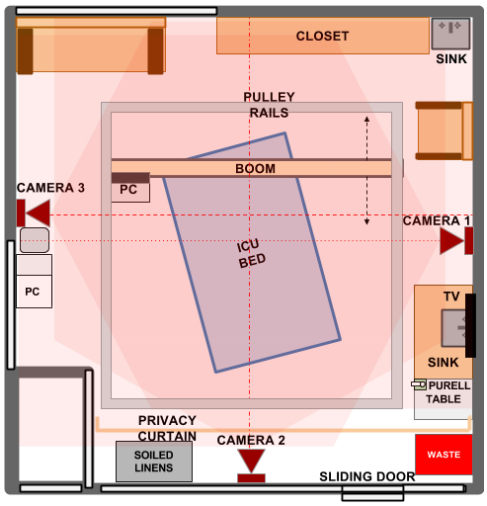


Figure 6.2: Top-view of the medical ICU.
The space is monitored by three nodes, each containing RGB-D sensors.

6.1.1 Importance of Event Logs and Qualifiers

Consider the issue of Hospital Acquired Infections (HAIs) by touch in the ICU. For consistency, assume that events can have one of four qualifiers: clean, contamination, transmission, or unclean. The labels depend on the sequence of underlying activities and the detection of hand sanitation activities performed by people after entering and before exiting the ICU room. The variation of HAI-relevant events is often attributed to staff fatigue, monotonous routines, or emergencies, and to visitors not being aware of sanitation protocols. The log's objective is to provide a chronological description of events inside the ICU room. The logs can be used by healthcare professionals to backtrack the origins of pathogens, which can help designing and executing corrective action plans.

Event and activity ICU logs will enable the following tasks for healthcare:

- An unobtrusive monitoring system for healthcare that can be used to detect deficiencies and areas of improvement, while maintaining the privacy of patients and staff.
- Logs to help create sanitation-corrective action plans.
- Semantic healthcare logs that can be used to analyze the spread of pathogens and HAIs by touch and others.
- A platform to evaluate best practices in ICU architectural and operational designs, which promote sanitation and prevent the spread of infections.

6.1.2 Related Technical Work

The latest developments in convolutional neural network (CNN) architectures for visual activity recognition achieve impressive performance; however, these techniques require large labeled data sets [3, 9, 78, 79]. The method in [70] uses egocentric cameras to analyze off-center activities. The method in [81] uses CNNs to analyze off-center activities, but it requires scenes with good illumination and clear of occlusions. Multi-sensor and multi-camera systems and methods have been applied to smart environments [25,84]. The systems require alterations to existing infrastructure making their deployment in a hospital logistically difficult. Further, most existing methods are not designed to account for illumination variations and occlusions and do not account for non-sequential, subtle motion, thus limiting their applicability in typical ICU conditions.

Healthcare applications of patient monitoring include the detection and classification of patient body configurations for quality of sleep, bedsore incidence, and rehabilitation. In [75], the authors introduce a coupled-constrained optimization technique that allows them to trust sensor sources for static pose classification. In [76], the authors use a mul-

timodal multiview system and combine it with time-series analysis to summarize patient motion. A pose detection and tracking system for rehabilitation is proposed in [52]. The controlled study in [55] focuses on workflow analysis by observing surgeons in a mock-up operating room. The work most similar to HEAL is introduced in [37], where Radio Frequency Identification Devices (RFIDs) and a single depth camera are used to analyze work flows in a Neo-Natal ICU (NICU) environment. These studies focus on staff activities and disregard patient motion. *Literature searches indicate that HEAL is the first of its kind in utilizing a distributed multimodal camera network for activity monitoring in a real hospital environment.* HEAL's technical innovation is motivated by medical needs and the availability of cheap sensors and ubiquitous computing. It observes the environment and extracts contextual aspects from various ICU room activities. The events are observed from multiple views and modalities. HEAL integrates contextual aspects such as roles and interactions with temporal information via elastic-net optimization and principled statistics.

A sample input and output for activity classification is shown in Figure 6.3, where various activity elements are identified across the multiviews of the ICU.

Figure 6.3: HEAL views of the mock-up ICU
The view shows where HEAL is tested and activities and events are simulated with the help of actor-volunteers. Left: the multiview (depth and grid information are not shown) input videos for HEAL. Right: the labeled activity output and its labeled aspects.

Technical Contributions. The main contributions are:

- A framework that defines and clearly integrates contextual aspects to identify activities and create event logs.
- The concept of role and role identification, which narrows the activity-role label search space.
- The ability to detect and analyze concurrent activities based on space-time constraints.
- Integration of activity regions (activity location and interaction cones) to further narrow the activity label search-space and improve activity detection and classification.
- An activity duration modeling technique that discriminates activities based on their duration.

6.2 HEAL Events and Activities Dataset

Two experimental setups are considered. First, we built a mock-ICU room complete with an ICU bed and various activities are acted out. The multimodal sensor rig was custom built as described below using off-the-shelf components and Raspberry Pi3 devices for data acquisition. This provided the preliminary data for methods development.The mock-up data contains 30-minute videos from six views, each view having two modalities. The videos are fully annotated. The preliminary data enabled us to modify the acquisition process and deploy a fully functional distributed sensor network in a community hospital.

System Setup. The sensor network is composed of three independent nodes each with a RaspberryPi 3B+, an RGB-Depth carmine camera sensor, and a battery pack. The

elements are placed inside an aluminum enclosure for sanitation purposes. The nodes are placed at three distinct locations in the ICU to ensure complete space coverage as shown in Figure 6.2. The nodes use TC/IP protocols for communication and synchronization via a Local Area Network. Each node can operate for up to 12hrs on a single battery.

Activity Set The 20 activities in the set α with their corresponding number of observed instances are: washing hands (68), sanitizing hands (33), entering the room (200), exiting the room (185), delivering food (15), delivering medicine (10), auscultating (48), cleaning room areas (16), cleaning the patient (18), bedside sitting (80), watching tv (45), patient moving on bed (50), rotating (adjusting) the patient (76), observing equipment (105), visiting patient without contact (83), visiting patient (with contact) (59), eating (16), sleeping/resting (84), turning lights on (60), turning lights off (45).

Event Set. This study covers the following set of events $\mathcal{E}$:

1. Clean: As people walk into the ICU room, they use hand sanitizers or wash their hands. After performing a series of activities, the person uses the hand sanitation once again, as the last activity before stepping out of the room.

2. Contamination: Occurs when visitors bring in contaminants or pathogens from outside the room by bringing in contaminated equipment, objects, or contaminated hands (unwashed or unsanitized).

3. Transmission: Occurs when an individual, such as a nurse, enters a room and follows sanitation protocols up to the point before leaving the room, bringing out contaminants and pathogens, which can affect others.

4. Unclean (risk of contamination and transmission): Occurs when sanitation protocols are not followed, neither upon entering nor exiting the room.

Figure 6.4 shows a sample log with sanitation event qualifiers. For instance, a very descriptive "clean visit event" includes the following sequence of activities: visitors enter the room, visitors sanitize their hands, visitors seat by the bed, visitors gets up, visitors sanitizes their hands one last time, and visitors exit the room. In short, the event is qualified based on hand sanitation and washing activities in the ICU immediately after entering and before exiting the room. HEAL only observes the inside of the room and not the outside, where additional sanitation and washing stations are also available.

When	Who	Entry Sanitation	What	Where	Exit Sanitation	Event Qualifier	Event Description
t_1	Visitor 1	YES	Visit	Chair	NO	Transmission	Sat, tablet, no contact
t_2	Doctor	YES	Check	Patient	YES	Clean	Auscultation, contact
t_3	Visitor 2	NO	Contact	By bed	YES	Contamination	Personal visit, contact
t_4	Visitor 3	NO	Visit	By bed	NO	Unclean	Stood, no contact
...	...	...	...	...	...	...	...
t_{T-2}	Unknown	NO	Janitorial	Room	NO	Unclean	Empty trash, no contact
t_{T-1}	Assistant	NO	Delivery	Bed	YES	Contamination	Delivered meds, contact
t_T	Caterer	YES	Exit	Room	NO	Transmission	Delivered food, no contact

Figure 6.4: Sample HEAL log.
The log records: time, role, activity label, location, and detailed description given by a human observer.

The following tasks are performed to classify activities: (1) detect people, (2) identify relevant objects, (3) define the activity blocks (location of respective activities), (4) estimate interaction cones from quantized poselet orientations, (5) estimate activity duration at the estimated ICU location from the grid, and (6) infer person roles. Tasks (1) to (4) are the activity and interaction regions, Task (5) is activity duration using HSMMs, and task (6) is achieved using interaction maps and allows HEAL to narrow the activity search space and increase its activity and event classification accuracy. Events

qualifiers are estimated from a sequence of activities, where the objective is to identify sanitation activities and localize them in time as immediately after entering or before exiting the ICU room.

6.3 Approach

The problem of event logging involves identifying *what* activities are executed, *where* these activities are executed, and by *whom*, in chronological order. In addition, interacting objects and the activity duration are also recorded. For example, consider the hand-washing activity: this involves a person (nurse) walking towards the sink, using the soap, drying with a towel, and walking away from the sink. The interacting objects are the sink, soap, and towel. The description includes the location of the sink and duration of the overall activity, the objects present, the locations where the person moved around within the monitored space, and the duration or time spent at these various locations. These cues provide significant contextual information is used to identify individuals and their activities. We refer to these data as *Contextual Activity Aspects*, represented by a P-dimensional vector described as follows.

6.3.1 Contextual Activity Aspects

Contextual aspects capture the location, orientation and interaction of a person with other static and dynamic objects in the ICU scene. We use two major objects categories: tagged (i.e., initialized manually: patient, patient bed, sink, ventilator, etc.) and automatically detected (e.g., cart, cot bed, bottles, books, other people). The P aspects ($P = 327$ in our implementation) are 40 Interaction Cones (10 tagged Objects $\times$4 orientations) + 3 Duration levels + 256 Grid Blocks + 20 detected Objects + 8 Roles.

1. Interaction Cones (C). The interaction cone vector is a vector with 4× number-of-tagged-ICU-objects elements. Its elements can take values from the set $\{1(close), 2(nearby), 3(far)\}$ depending on the distance to the object. In our implementation we use 1: $\leq$ 1ft, 2: $>$ 1 $\leq$ 2ft, 3: $>$ 2ft). The cone vector ($\mathbf{f}_{cone}$) encodes relative orientation and distance to objects of interest. There are 10 identified ICU objects (light-switch, bed, ventilator, trashcan, computer, closet, couch, door, sink, and tv), so the cone vector has 40 elements.

2. Activity Grid (G). The monitored space is partitioned into a Cartesian map with $G = g \times g$, where g is the grid dimensions. The map encodes activity location as a G dimensional binary vector $\mathbf{f}_{grid}$. Our implementation uses a 16×16 grid, so $G = 256$.

3. Activity Duration (D). The activity duration is modeled using segments to more flexibly account for variable state longevity and quantized into slow, medium, and fast. The duration vector is represented by $\mathbf{f}_{duration}$.

4. Foreign Objects (O). The of detectable objects include: laptops, trays, chairs, carts, boxes, cups, books, etc. In our implementation, the number of detectable objects per activity is limited to max of 20. This vector is represented by $\mathbf{f}_{objects}$.

5. Roles (R). Eight actor roles are considered – nurse Assistant, Caterer, medical Doctor, Facilities, Isolation, Nurse, Patient, and Visitor [74]. The role aspect is an eight-element vector, where each element is the score assigned to the corresponding role. The role vector is represented by $\mathbf{f}_{role}$.

6.3.2 Activity Representation via Contextual Aspects

Let $\mathbf{f}_{aspects} = [\mathbf{f}_{cone}, \mathbf{f}_{map}, \mathbf{f}_{duration}, \mathbf{f}_{objects}, \mathbf{f}_{roles}]$, $\mathbf{f}_{aspects} \in \mathbb{R}^P$, represent the contextual aspects feature vector computed at each frame n (the frame number is omitted to simplify the notation) for *each detected person* in the scene. Ideally one could use this aspects vector for the modeling and recognition stages. However, given the uncertainty and noise in the measurements, we found that it is more effective to perform this analysis after approximating the vector in a reduced basis representation. This approximation $\mathbf{f}^{(M)}$ is composed as a linear combination of M aspects basis $\mathbf{\Phi} = \big[\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_M\big], \phi_m \in P$ and $M \leq P$ reconstruction weights $\mathbf{w} = \{w_1, \ldots, w_M\}, 1 \leq m \leq M$, with

$$\mathbf{f}^{(M)} = \sum_{m=1}^{M} w_m \boldsymbol{\phi}_m. \tag{6.1}$$

6.3.3 Aspect Basis and Weights

The aspect basis and aspect weights are estimated from the collection of K segmented and labeled training frames by minimizing:

$$\underset{\mathbf{\Phi}, \mathbb{W} \in \mathbb{R}^{\mathbf{M} \times \mathbf{N}}}{\text{minimize}} \sum_{n=1}^{N} \left(\frac{1}{2} \|\mathbf{f}_{aspects_n} - \mathbf{\Phi}\mathbf{w}_n\|_2^2 + \gamma \|\mathbf{w}_n\|_1 \right) \tag{6.2}$$

where $\mathbb{W} = [\mathbf{w}_1, \ldots, \mathbf{w}_N] \in \mathbb{R}^{P \times N}$, γ $(= 0.2)$ is the regularization parameter. The solution to Eqn.(6.2) is coded in Python using the optimization library from [11].

6.4 Aspects Computation

Contextual aspects are computed *for each detected person per time instance* in the scene. Their computation involves the following steps: tag static object of interest,

detect individuals entering the room, compute appearance features and initialize a depth-modality blob tracker, estimate poselets and compute interaction cones, detect foreign objects, and estimate roles. Multiple person detectors are tested and two are selected for system deployment [10] limited by Raspberry Pi hardware and convolutional neural networks [63] for offline analysis. This section describes the aspects computation: cones, duration, grid, objects, and roles.

6.4.1 Interaction Cones Aspect

Individuals are tracked using the Depth modality via a blob tracker and RGB modality using the method from [51]. The location of individuals is mapped between RGB and Depth modalities and localized on the activity grid map. Finally, the poselet detector from [6] is used to estimate the relative pose orientation of a person with respect to the door-way. The orientation is quantified using a conical structure shown in Figure 6.5. A cone is one of four circumference quadrants. Each cone has a 90^o operating arc starting at the 315^o mark. The elements of the cone feature vector $\mathbf{f}_{cone} = \{C_{o,q}\}, 1 \leq o \leq O, 1 \leq q \leq 4$ contain the distances ($C_{o,q} = d$) between the individuals and each object o from the set of tagged ICU objects $\mathcal{O} = \{$bed, chair, computer, doorway, nearest-person, sink, table, trashcan, closet, and ventilator$\}$.

6.4.2 Activity Duration Aspect

A major limitation of existing activity recognition and classification methods is the inability to distinguish activities that appear to be similar, i.e., coming from a similar scene context. For example, in the ICU environment, walking by the sink, sanitizing hands, and washing hands all appear very similar. The challenge is to identify the aspects that provide discriminant feature representations of these activities. We use the HSMM

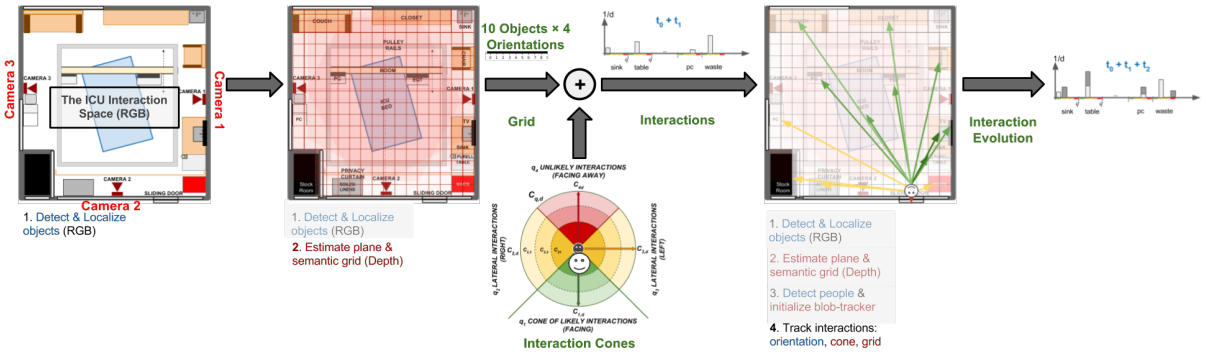


Figure 6.5: Simplified Interaction Process.
The interaction cones represents relative orientation and distances between individuals and tagged objects of interest to the ICU.

from [60] as it offers a flexible modelling of activities duration as opposed to conventional HMM. Figure 6.6 shows the modified trellis and its components. Our implementation uses the software library from [31]. The sequence of states $y_{1:T}$ is represented by the segments ($\mathbf{\Omega}$). A segment is a sequence of unique, sequentially repeated observations (person grid locations). The segments contain information to identify when the person is detected, what the person is doing, and for how long (in time-slice counts). The elements of the j-th segment (Ω_j) are the indices (from the original sequence of locations) where the observation (b_j) is detected, the number of sequential observations of the same symbol (duration d_j), and the state or pose (y_j).

HSMM elements. The hidden variables are segments $\Omega_{1:U}$ and the observable features are $X_{1:T}$, which are the semantic grid vectors. The joint probability of the segments and the semantic activity location features is given by:

$$\Pr(\Omega, X) = \Pr\left(\Omega_{1:U}, X_{1:T}\right) = \Pr\left(Y_{1:U}, b_{1:U}, d_{1:U}, X_{1:T}\right) \tag{6.3}$$

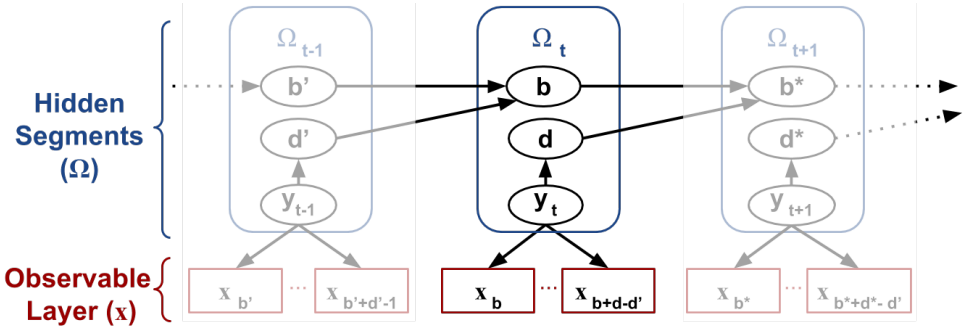


Figure 6.6: HSMM trellis.
The elements are: hidden segments Ω_j indexed by j and their elements $\{b_j, d_j, y_j\}$. The variable b is the first detection in a sequence, y is the hidden layer, (x) is the observable layer containing samples from time b to $b + d - d'$. The observation's initial detection and observation's duration are b and d, respectively.

$$\begin{aligned}\Pr(\Omega, X) = &\Pr(y_1)\Pr(b_1)\Pr(d_1|y_1) \times \prod_{t=b_1}^{b_1+d_1+1} \Pr(x_t|y_1) \\ &\times \prod_{u=2}^{U} \Pr(y_u|y_{u-1}) \Pr\left(b_u|b_{u-1}, d_{u-1}\right) \\ &\times \Pr\left(d_u|y_u\right) \prod_{t=b_u}^{b_1+d_1+1} \Pr(x_t|y_u),\end{aligned} \tag{6.4}$$

where U is the sequence of segments such that $\Omega_{1:U} = \{\Omega_1, \Omega_2, ..., \Omega_U\}$ for $\Omega_u = \left(b_u, d_u, y_u\right)$ and with b_u as the start position (a bookkeeping variable to track the starting point of a segment), d_u is the duration, and y_u is the hidden state ($\in \{1, ..., Q\}$). The range of time slices starting at b_u and ending at $b_u + d_u$ (exclusively) have state label y_u. All segments have a positive duration, a time-span $1 : T$ without overlap, and are constrained by:

$$b_1 = 1; \qquad \sum_{u=1}^{U} d_u = T; \quad \text{and} \quad b_{u+1} = b_u + d_u. \tag{6.5}$$

The transition matrix ($\mathbf{\Psi}$): $\Pr(y_u|y_{u-1})$, represents the probability of going from one

segment to the next via:

$$\mathbf{\Psi} : \Pr\big(y_u = j | y_{u-1} = i\big) \equiv \psi_{ij} \tag{6.6}$$

The first segment (b_u) starts at 1 ($u = 1$) and consecutive points are calculated from the previous point via:

$$\Pr\big(b_u = \beta | b_{u-1} = \nu, d_{u-1} = l\big) \text{ is } \delta\big(\beta - \nu - l\big) \tag{6.7}$$

where $\delta(i - j)$ is $1 : i = j; 0 :$ else. The dummies are $\beta = \nu + l$, with β, ν, l, and $i = j$.

Finally, the probability of duration d_u is given by:

$$\Pr(d_u = l | y_u = i) = \Pr_i(l) \tag{6.8}$$

Using segments and HSMMs we can model the state duration as a normal distribution $\Pr_i(l) = \mathcal{N}_{l,i}(\mu, \sigma)$ and the duration probability of the i-th state can be used to distinguish between slow, medium, and fast activities. We refer to Chapter 4 and the work in [76] for details about HSMM parameter estimation and inference processes.

Activity duration is analyzed at three levels: slow, medium, and fast. This allows us to further reduce the label search space. For example, duration information is used to distinguish washing hands (slowest), sanitizing hands (moderate), or walking by the sanitation station (fastest) activities. Additional aspects such as detected objects, critical object interactions, activity locations, and person roles is extracted from the training videos to increase the probability of correctly identifying activities and logging events.

6.4.3 Activity Grid Aspect

The binary grid vector $\mathbf{f}_{grid} = [g_1, \ldots, g_{16}, \ldots, g_{256}]$ represents activated activity regions and are computed per person. The spatial location is computed by overlaying a 2-D grid on the ICU work-space as shown in Figure 5.1. The grid dimensions depend on the size of the physical space. When projected to the ICU floor, each block in the grid has dimensions 18×18 inches. The floor plane is estimated from three points using standard image geometry methods. The grid dimensions are $g \times g$ dimension with $g = 16$ yields a 256 element activity grid vector (i.e., $|\mathbf{f}_{grid}| = g \times g = 256$). A sample food delivery map is shown in Figure 6.7 overlayed in translucent black, indicating the areas where activities occur.

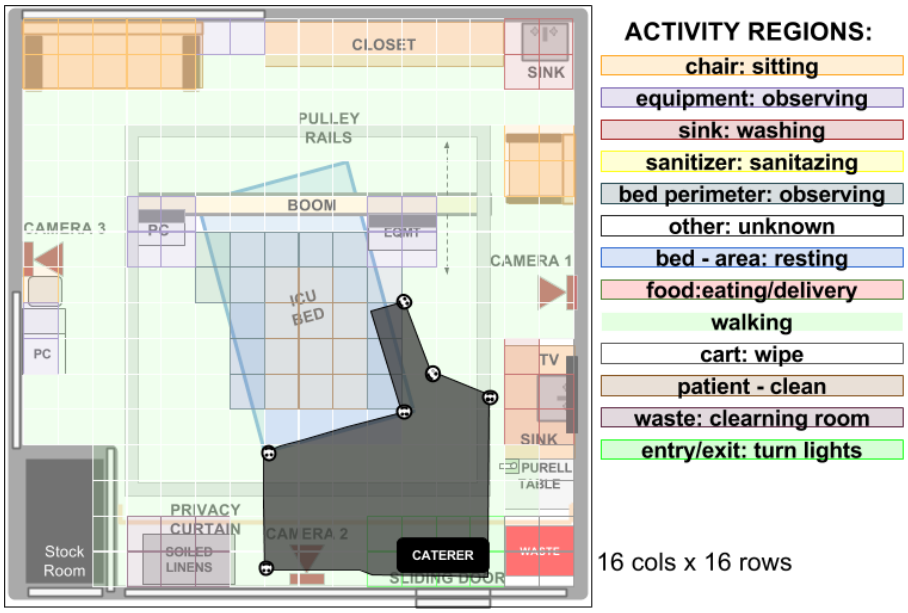


Figure 6.7: Caterer semantic activity map overlayed in black.
The grid dimensions are 16×16 and the various block colors represent activity regions and are described by the legend on the top-right corner.

6.4.4 Foreign Objects Aspect

Methods to detect foreign ICU object are tested. These include: [14], [36] and, [63]. There is an uncountable number of objects associated with activities. A total of 20 objects is selected based on a detection consistency $\geq 75\%$ on 10 continuous observations. Evaluation of object detectors is beyond the scope of this work; however, the best performing detector for offline-ICU processes is YOLO [63], which uses convolutional neural networks. The best detector running on the RPi3 is [14], which uses learned attributes.

6.4.5 Roles Aspect

Use of identifiable information in the ICU is restricted by patient privacy, labor protection, and Health Insurance Portability and Accountability Act (HIPAA) stipulations [16]. We use role representation from appearance and interaction information to deal with these ICU restrictions. It assigns roles over the complete activity or event using a threshold (70%) based on the number of frames or observations to link a role, else the role is considered to be "unknown". Learning a role starts with identifying appearance and interaction features for each role and compute scores for each element in the vector $\mathbf{f}_{role} = \{S_r\}_R$ for roles in the set $\mathcal{R}$ = [Assistant, Caterer, Doctor, Facilities, Isolation, Nurse, Patient, and Visitor], indexed by $r, 1 \leq r \leq R$, from all views $v, 1 \leq v \leq V$, and across all frames $n, 0 \leq n \leq N$. The appearance vectors ($\lambda_{n,r,v}$) are computed at $n = 0$ and used to construct the dictionary of appearances for all roles $\mathbf{\Lambda} = \{\Lambda_r\}_R$. Similarly, the interaction vectors ($\zeta_{n,r,v}$) are computed for $1 \leq n \leq N$ and are used to construct a dictionary of role-interactions for all roles $\mathbf{Z} = \{\mathrm{Z}\}_R$. The candidate scores are computed using the method described in Chapter 5 and [74] as $S_r, 1 \leq r \leq R$, which is a combination of appearance and interaction scores given by:

$$S_r = (S_r^{\Lambda} + S_r^{\mathrm{Z}}) \tag{6.9}$$

The estimated role R^* is the one with the most similar representation via:

$$R^* = \arg \min_{1 \leq r \leq R} (S_r) \tag{6.10}$$

6.5 Testing Contextual Activity Aspects

Activity Classification. Activity labels are estimated using the computed aspects basis and weights over an observed event with N frames indexed by $n, 0 \leq n \leq N$. Activity label inference is integrated via majority-vote over the range of frames that starts at frame n_i and ends at frame $n_o, 0 < n_i \leq n_i + h$ and $n_i + h = n_o \leq N$. Activity labels are estimated using the per-frame aspect information, where h is size of activity-observation window in number of frames. Our implementation uses $h = 6$ (approximately 1 second). Activity label scores S_a are obtained via:

$$S_a = \sum_{n=n_i}^{n_o=n_i+h} \mathrm{D}(\mathbf{f}_n, \theta_a), 1 \leq n_i \leq N - h, \tag{6.11}$$

where $\mathbf{f}_n$ is defined in Eqn. (6.1), with its elements computed using Eqn. (6.2), where θ_a is the LDA-decision hyper-plane of activity $a \in \alpha$. Finally, the activity label a^* is:

$$a^* = \arg \max_{a \in \alpha} (S_a). \tag{6.12}$$

Ambiguous activities are labeled unknown and identified via the ratio test on S_a with a relative dissimilarity of, at least, 0.2 for the highest and second-highest label candidates. Due to the limited number of instances a very small number of activities, such as clerical, physical therapy sessions, and religious services, are classified as unknown.

6.5.1 Event Log Creation

A sample log in shown in Figure 6.4. The various aspects are used to populated the log as shown in Figure 6.1 and described as follows: First, the ICU door is used to mark

the beginning frame ($n = 0$) and the ending of an event ($n = N$); single individuals are detected, tracked, and localized using the grid map and the blob-tracker; finally interactions, activity duration, and role aspects are computed to infer activity labels for a set of frames starting at $n_i, i \geq 1$ and ending at $n_o, o \leq N$; The activities are localized in time using the order of the frames and combined with the aspect values to populate the log. The event qualifiers are estimated after the conclusion of the vent (i.e., individual exited the room).

Event Qualifier Estimation. Consider the event E given by a sequence of J activities indexed by j i.e., $E = [a_1, \ldots a_J] = \{a_j\}_J,\ 1 \leq j \leq J$. A single activity is represented by a. The event qualifiers evaluate the order of sanitation (hand-washing or hand sanitation) activities in a sequence of activities and provide sanitation labels based on detection window at the beginning and at the end of a sequence. In our implementation, we consider a clean entry if sanitation is detected within the first three activities. Similarly, a clean label is assigned if sanitation activity is detected within the last three activities.

6.6 HEAL Experimental Results

HEAL is evaluated using a 10-fold cross-validation. The reported results are the confusion matrix obtained from the best fold and the mean accuracy over all folds. Figure 6.8 shows the affect of M on activity classification accuracy.

Accuracy of Log Event Qualifiers. Logs are descriptions of past events that occurred in an area and were performed by a certain role. This experiment involves evaluating the correctness of the event qualifiers: clean, contamination, transmission, and unclean qualifiers by asserting that a sanitation event is detected within the first activities performed by an individual that entered the room and within the last three activities

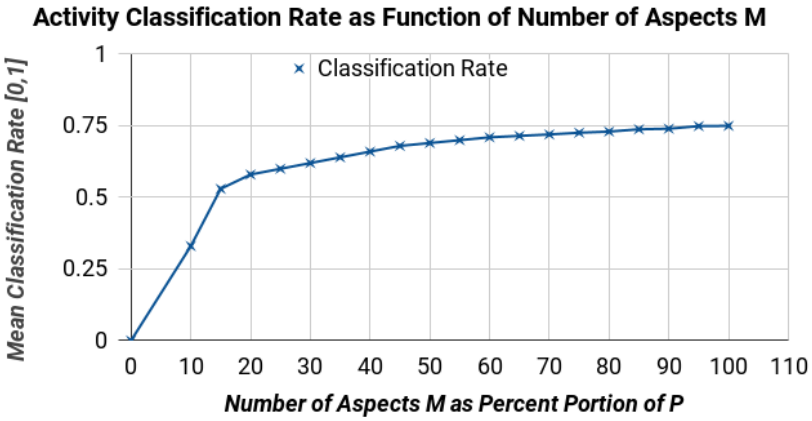


Figure 6.8: Mean activity classification accuracy as function of M aspects. The M aspect weights are $\{w_m\}_M, M \leq P$ for basis $\{\phi_m\}_M, \phi_m \in P$.

performed by an individual that exited the room. The confusion matrix in Figure 6.9 shows true and predicted event qualifiers, with an 82.5% average accuracy rate.

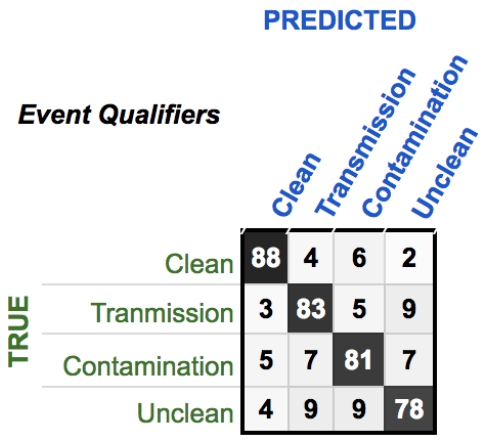


Figure 6.9: Confusion matrix for the ICU sanitation-event qualifier estimation. The cells are color scaled to indicate accuracy (darker cells correspond to higher classification accuracy) in scale 0-100.

Contribution of Aspects for Activity Classification. This experiment demonstrates the impact of the contextual aspects in activity classification.

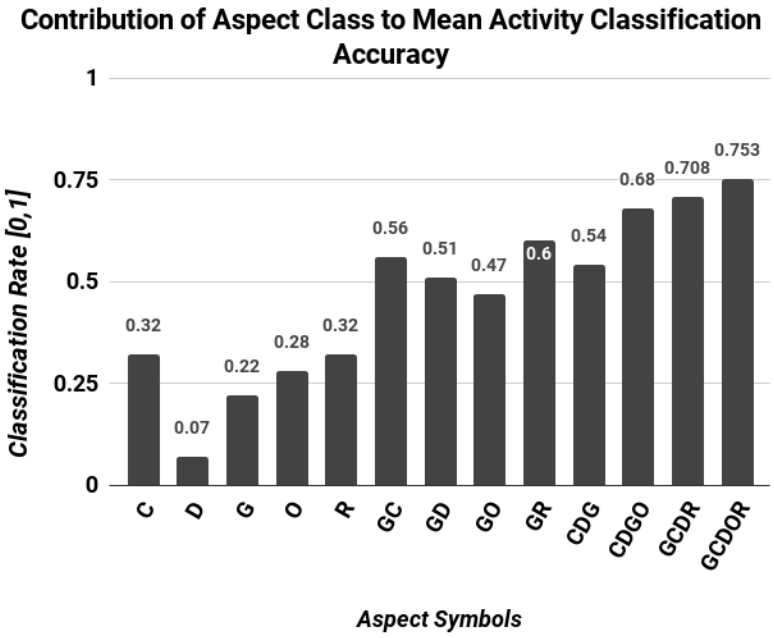


Figure 6.10: Contribution of contextual aspects for mean classification accuracy. The aspects are: the interaction cones (C), the activity duration (D), the activity grid location (G), the object detector outputs (O), and the roles (R).

Activity Classification. Even with the use of aspects activities in the ICU can be confused with other, similar activities. The confusion matrix in Figure 6.11 shows the labels and rates of correctly and incorrectly classified activities, where darker cells correspond to better performance and the rows add up to 100. The left column contains the true labels and the top row the predicted labels.

The bar-plot in Figure 6.12 compares the proposed approach to two methods: the in-house implementation of [37], which classifies activities using RFIDs and a single depth camera via distance feature vectors and a support vector machine (SVM); and [78], which uses C3D features with a linear SVM. In [37] the authors use distances to represent person-object interactions for healthcare staff. However, it does not include interactions, roles, or activity duration. The C3D method uses deep convolutional operations, which

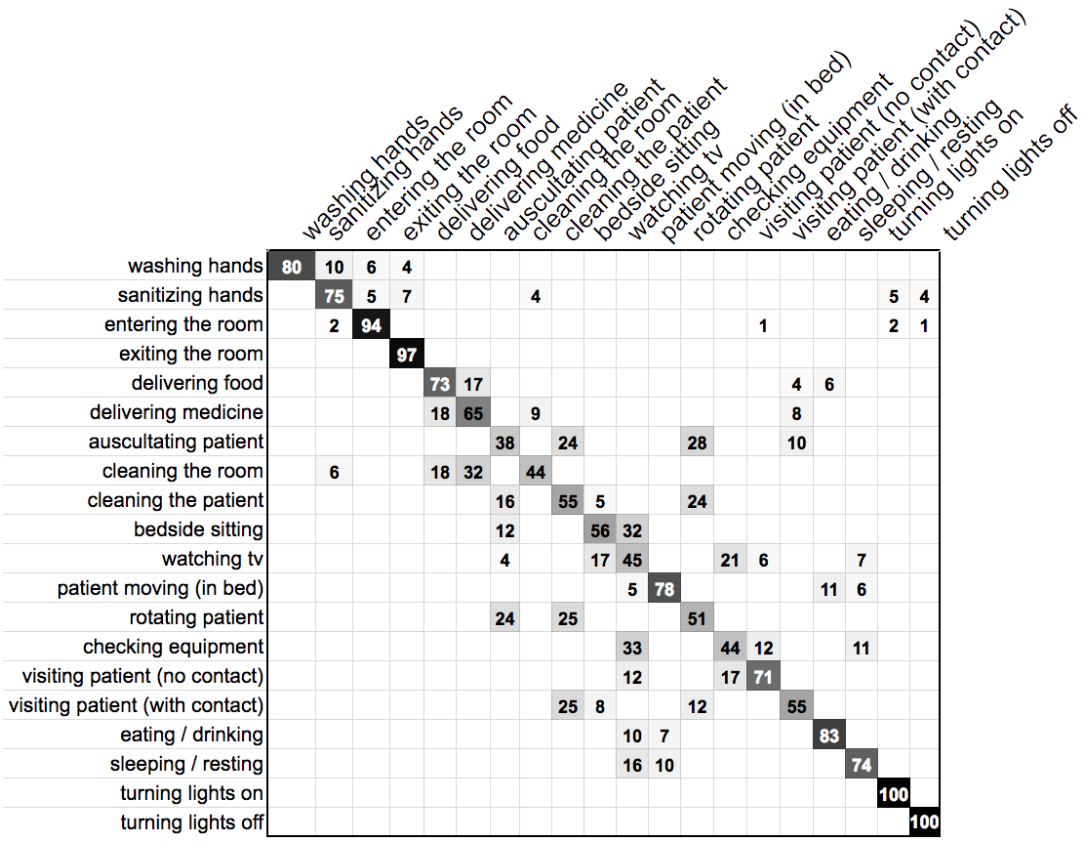


Figure 6.11: Activity classification confusion matrix of HEAL.
The left column indicates the true activity labels, while the top row (vertical text) indicates predicted activity labels. Darker cells indicate better performance, while empty cells indicate zero. The values are rounded and a [0-100] scale.

are unable to capture activities' contextual information. Neither of these methods encapsulates the subtleties captured by the contextual aspects such as activity regions, interactions, roles, and relative distances and orientations. This information helps to better represent and classify complex activities and allows the proposed solution to outperform the competition. The contextual aspects and their respective contribution for activity classification are shown in Figure 6.10. HEAL outperforms [37] by mean average classification ranging from 0.01 in "delivering medicine" to 0.31 in "sleeping/resting". The performance comparison between HEAL and C3D ranges from C3D outperforming

HEAL by 0.05 for "bedside sitting" to HEAL outperforming C3D in all other activities ranging from 0.1 for "exiting room" activity to 0.5 for "washing hands" activity.

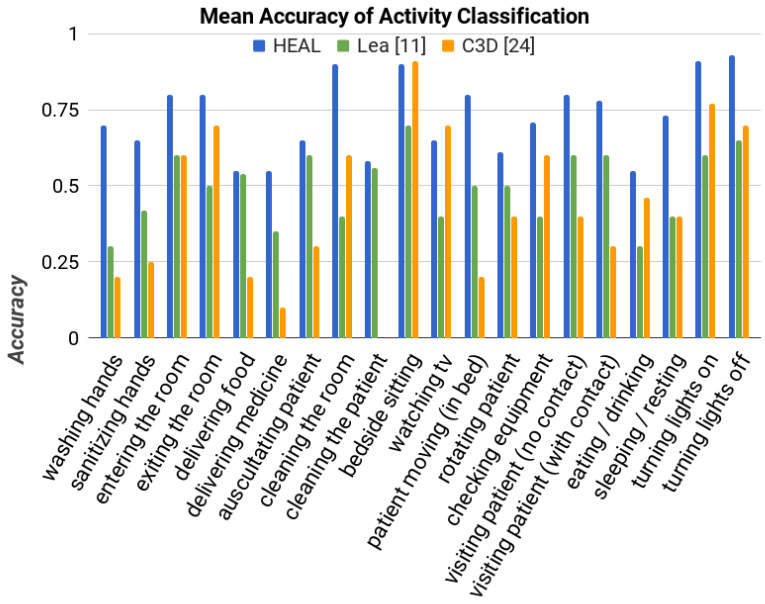


Figure 6.12: Classification accuracy of HEAL compared to the in-house implementation of two competing methods.

The work in [37] analyzes activities in a neo-natal ICU room the CNN method from [78], which is a popular technique to represent and classify activities.

6.7 Summary

We proposed a comprehensive multiview multimodal framework for robustly estimating sanitation qualifiers for events in an ICU. This is achieved by effectively leveraging contextual aspect information. The experimental results indicate that aspects contribute differently to the representation and classification of activities, estimation of event qualifiers, and creation of event logs. The methods rely on effective localization of individuals and objects in the ICU. The strength of HEAL is its multimodal multiview nature, which

allow the methods to robustly and effectively represent activities by detecting, tracking, and localizing person-objects and person-person interactions in the ICU.

HEAL is capable of robustly estimating the qualifiers for events in the ICU (i.e., clean, contamination, transmission, and unclean events). This is possible based on the relevance of the contextual aspects. However, the complexity and diversity of the actions in the ICU continue to be challenging problems.

Possible Future Direction. Open problems include explore applications outside the ICU such as supporting elderly independent living and monitoring and presenting effective logs for concurrent activities and events. The experimental results indicate that the various aspects contribute differently to the detection of actions, the estimation of event qualifiers, and proper event logging. A potential future direction includes presenting effective logs for concurrent actions and events. This future investigation may incorporate user studies to identify the best practices for presenting logs to medical practitioners.

Future opportunities can focus on the development and evaluation of new methods to analyze activities and events using Convolutional Neural Nets including training and re-evaluating the C3D network. Finally, future directions may analyze concurrent activities: multiple people performing multiple activities and multitasking: single individuals performing multiple activities.

Chapter 7

Discussion

This Dissertation introduced multiple concepts, algorithms and techniques to monitor patients and staff in healthcare environments. The proposed infrastructure is nonintrusive, non-disruptive and works without modifying the existing hospital infrastructure. The methods are modular, comply with hospital privacy protection stipulations, and robustly monitor the environment. The selected applications include sensor network design and evaluation; patient pose classification with multiple sensors and multiple modalities and views; patient pose pattern analysis from multimodal multiview data; healthcare role representation and identification; and healthcare activity and event analysis and logging.

Network elements and sensors are selected and evaluated for patient sleep pose classification in Chapter 3. This chapter describes how the network is further developed to remove sensors that required contact with patients. The evolution made the network a purely visual system, which uses multimodal mutiview data to accurately classify poses. Also, this chapter tackles modality and view trust estimation, while improving classification of poses. The evaluations include network configurations with incomplete data (e.g., simulation of sensor occlusions and malfunctions). The applications focus on static

detection and classification of patient sleep poses in a mock-up and a medical ICU over a large range of simulated and natural scene conditions, which include variable levels of illumination and occlusions.

The methods for patient sleep pose pattern analysis are introduced and evaluated in Chapter 4. This Chapter explores a dynamic analysis of patient poses. Poses are represented using deep features, which improve static classification accuracy. Pose transitions are represented and compressed using a proposed multimodal multimodal keyframe extraction algorithm. Pose transition patterns are analyzed using Hidden Markov Modeling (i.e., sequences of pseudo-poses). Finally, the pose patterns differences between poses and pose-transitions (i.e., pseudo poses) is analyzed using Hidden Semi-Markov Models (HSMM). The proposed HSMM-based technique uses variable duration segments of time, instead of uniform time samples from conventional HMM. HSMM is adapted to model pose and pseudo pose duration and quantify patient rate and range of motion.

In Chapter 5, the dynamic study is expanded from the patient to include all individuals in the ICU. However, there are limitations and restrictions that prohibit the use of identifiable information (i.e., HIPAA). Role representation and identification from combined appearance and semantic interaction maps is introduced to circumvent the use of identifiable information and to minimize the limitations of appearance based identification methods to natural scenes.

The last portion of this dissertation introduced the concept of contextual aspects bases and weights for activity and event analysis in Chapter 6. This last chapter integrates the findings regarding interactions, roles, objects, and flexible activity duration modeling from the previous chapters into a comprehensive set of recognition and analysis tasks.

7.1 Future Directions

This thesis introduced multiple problems and proposed solutions and methods that use a non-intrusive non-disruptive multimodal multiview distributed network of sensors. Nevertheless, a sea of problems remain unsolved and questions unanswered. For instance, A method for multimodal multiview trust estimation was introduced and successfully implemented, the method needs to be expanded from image classification to video analysis.

In pose pattern analysis methods are introduced to represent pose transitions, to extract keyframes, and to model pose duration and distinguish between poses and pseudo-poses. However, the proposed technique uses complete data and needs to be further develop to consider modality interdependence and data correlations. This analysis is essential so that solutions can infer pose patterns even with incomplete data and to effectively deal with sensor failures. Privacy restriction limitations are handled using role representations and identification. The roles are represented as a combination of appearance and interaction maps. However, the maps are computed using dynamic scenes. Although object detectors are used in a semi-automated fashion, effective and scalable solutions requires that the proposed techniques be expanded to deal with dynamic scenes or to further developed to combine detected activities (e.g., auscultating patient to indicate nurse or doctor) with observed interactions (e.g., relative locations and orientations visited by nurses or doctors) to infer roles.

Additional future direction includes integrating other modalities that can provide feedback such as thermography, which can be used to validate contact. Last but not least, questions regarding medical applications, clinical validation and feasibility of the proposed solutions methods, and algorithms remains as open questions, which can only be really answered by medical and clinical researchers.

Appendix A

Ubuntu Mate 16.04 on Raspberry Pi3 B

The Raspberry Pi3 B+ (RPi3) runs its operating system from an SD card. In this appendix we provide detailed steps to install Ubuntu Mate 16.04 on the RPi3 with an ARM 7 Quad Core processor. Detailed installation instructions regarding data collection libraries and drivers can be found in the MICU public repository.

A.1 Prepare the Elements

Using an Ubuntu do the following:

- Format the SD card using the fat32 format (recommend using a 32 Gb card)
- Download the 16.04 Ubuntu Mate image from ubuntu-mate to an Ubuntu host computer.

A.2 Install Ubuntu - Steps Executed on Host

- Press "alt" key and type/search for "Disks".
- Insert the SD card.
- Accept and execute.
- Once finalized, power-off the "Disks" application and remove the SDcard.
- Insert the card in the RPi3 device, connect the peripherals, and power on the RPi3.
- Follow the setup and optimization instructions prompted on the RPi3 screen.

A.3 Install OpenCV

An installation script is provided at the link: install-opencv.sh

A.4 Install OpenNi2

The dependencies are libraries are installed using the bash script from install.sh

A.5 Network Communication

The devices use TCP/IP communication via a local area network using a server/client configuration. The Python implementation of the client (client.py) and the server (server.py) are also posted on a public Github repository.

Appendix B

Install Ubuntu 12.04 on Older OMAP Devices

The following instruction cover the installation of libraries and dependencies on the PandaBoard ES and the BeagleBoard-XM. These devices were used in early generation of the system used presented in this thesis. However, both devices have very limited support and were ultimately replaced by the more powerful Raspberry Pi 3B+ modules.

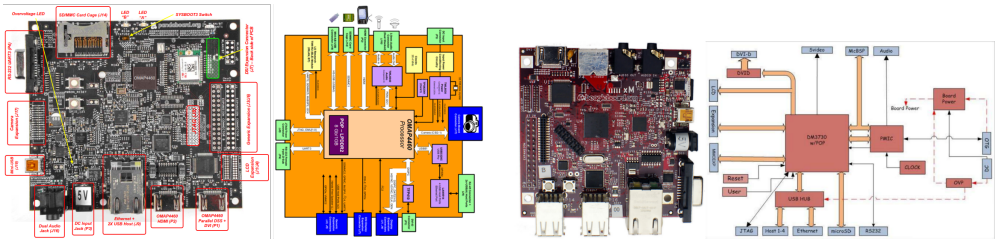
B.0.1 PandaBoard ES

The PandaBoard ES, shown in Figure B.1 (a), runs its operating system from an SD card. The following steps provide a detailed instructions to install Ubuntu 12.04 on the PandaBoard ES.

B.0.2 Prepare the Elements

Execute the following steps using an Ubuntu host:

1. Format the SD card using the ext4 format (recommend using a 32 Gb card)



(a) (b)

Figure B.1: ARM computers.
(a) PandaBoard ES ARM 8 Duo Processor., and (b) BeagleBoard-XM ARM 7 Processor. Both board were used in the preliminary healthcare environment deployment.

2. Download the 12.04 OMAP4 image to a known location from:

   Ubuntu release: http://cdimage.ubuntu.com/releases/12.04/release/

   Direct link: http://tinyurl.com/cpytu67

B.0.3 Install Ubuntu

3. Open a terminal (keyboard shortcut: ctrl + alt + t).

4. Check which devices are connected (without the SD card inserted).

   $ df -h

5. Insert the SD card and repeat (note the new device).

   $ df -h # new item >>sdd1 (in my case)

6. Unmount the detected SD card.

   $ sudo umount /dev/sdd1

7. Copy over the image to the SD card.

   $ cd Downloads $ gunzip -c ubuntu-12.04-preinstalled-desktop-armhf+omap4.img.gz

   — sudo dd bs=4M of=/dev/sdd

   # Note that the device is labeled **sdd1** but mounted on location **sdd**.

B.1 Install Ubuntu 12.04 on BeagleBoard-XM

The BeagleBoard-XM runs its operating system from an SD card. In this appendix we provide a detailed set of steps to install Ubuntu 12.04 on the BeagleBoard-XM with an ARM 7 Single Core processor.

This appendix covers the installation of Bubuntu 12.04 on the BeagleBoard-XM ARM7 computer.

B.1.1 Prepare the Elements

Using an Ubuntu do the following:

- Format the SD card using the fat32 format (recommend using a 32 Gb card)
- Download the 12.04 OMAP image from `www.google.com` to an Ubuntu host

B.1.2 Install Ubuntu

- Open a terminal (keyboard shortcut: ctrl + alt + t).
- Without the card inserted and check the devices connected (cmmd: df -h).
- Insert the card and repeat (cmmd: df -h). Note the new item.

Bibliography

[1] M. A. R. Ahad, J. K. Tan, H. Kim, and S. Ishikawa. Motion history image: its variants and applications. *Machine Vision and Applications*, 2012.

[2] B. B. Amor, J. Su, and A. Srivastava. Action recognition using rate-invariant analysis of skeletal shape trajectories. *IEEE transactions on pattern analysis and machine intelligence*, 38(1):1–13, 2016.

[3] M. Baccouche, F. Mamalet, C. Wolf, C. Garcia, and A. Baskurt. Sequential deep learning for human action recognition. In *International Workshop on Human Behavior Understanding*, pages 29–39. Springer, 2011.

[4] L. Bazzani, M. Cristani, and V. Murino. Symmetry-driven accumulation of local features for human characterization and re-identification. In *Elsevier Computer Vision and Image Understanding (CVIU)*, 2013.

[5] S. Bihari, R. D. McEvoy, E. Matheson, S. Kim, R. J. Woodman, and A. D. Bersten. Factors affecting sleep quality of patients in intensive care unit. *Journal of clinical sleep medicine: official publication of the American Academy of Sleep Medicine*, 8(3):301, 2012.

[6] L. Bourdev and J. Malik. Poselets: Body part detectors trained using 3d human pose annotations. In *Proc. of Int'l Conf. on Computer Vision (ICCV)*. IEEE, 2009.

[7] I. Brault, K. Kilpatrick, D. DAmour, D. Contandriopoulos, V. Chouinard, C.-A. Dubois, M. Perroux, and M.-D. Beaulieu. Role clarification processes for better integration of nurse practitioners into primary healthcare teams: a multiple-case study. In *Nursing research and practice*. Hindawi Publishing Corporation, 2014.

[8] D. S. Cheng, M. Cristani, M. Stoppa, L. Bazzani, and V. Murino. Custom pictorial structures for re-identification. In *British Machine Vision Conf. (BMVC)*, 2011.

[9] G. Chéron, I. Laptev, and C. Schmid. P-cnn: Pose-based cnn features for action recognition. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3218–3226, 2015.

[10] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2005.

[11] S. Diamond, E. Chu, and S. Boyd. CVXPY: A Python-embedded modeling language for convex optimization, version 0.2. `http://cvxpy.org/`, 2014.

[12] M. Ester, H.-P. Kriegel, J. Sander, X. Xu, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Kdd*, 1996.

[13] G. Evangelopoulos, A. Zlatintsi, A. Potamianos, P. Maragos, K. Rapantzikos, G. Skoumas, and Y. Avrithis. Multimodal saliency and fusion for movie summarization based on aural, visual, and textual attention. *IEEE Transactions on Multimedia*, 2013.

[14] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth. Describing objects by their attributes. In *IEEE Proc. of Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2009.

[15] G. Farnebäck. Two-frame motion estimation based on polynomial expansion. *Image analysis*, pages 363–370, 2003.

[16] C. for Medicare & Medicaid Services et al. The health insurance portability and accountability act of 1996 (hipaa). *Online at http://www. cms. hhs. gov/hipaa*, 1996.

[17] Y. Freund and R. E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*, 55(1):119–139, 1997.

[18] J. García, N. Martinel, A. Gardel, I. Bravo, G. L. Foresti, and C. Micheloni. Modeling feature distances by orientation driven classifiers for person re-identification. *Elsevier Journal of Visual Communication and Image Representation*, 2016.

[19] T. Giraud, J.-f. Dhainaut, J.-f. Vaxelaire, T. Joseph, D. Journois, G. Bleichner, J.-p. Sollet, S. Chevret, and J.-f. Monsallier. Iatrogenic complications in adult intensive care units: a prospective two-center study. *Critical care medicine*, 1993.

[20] S. J. Gordon, K. A. Grimmer, and P. Trott. Understanding sleep quality and waking cervico-thoracic symptoms. *Inet. Journal of Allied Health Sciences and Practice*, 2007.

[21] D. Gray and H. Tao. Viewpoint invariant pedestrian recognition with an ensemble of localized features. In *Springer Proc. of European Conf. on Computer Vision (ECCV)*, 2008.

[22] T. Grimm, M. Martinez, A. Benz, and R. Stiefelhagen. Sleep position classification from a depth camera using bed aligned maps. In *IEEE Int'l Conf. on Pattern Recognition (ICPR)*, 2016.

[23] C. Guérin, J. Reignier, J.-C. Richard, P. Beuret, A. Gacouin, T. Boulain, E. Mercier, M. Badet, A. Mercat, O. Baudin, et al. Prone positioning in severe acute respiratory distress syndrome. *New England Journal of Medicine*, 2013.

[24] R. I. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, ISBN: 0521540518, second edition, 2004.

[25] E. Hoque and J. Stankovic. Aalo: Activity recognition in smart homes using active learning in the presence of overlapped activities. In *2012 6th International Conference on Pervasive Computing Technologies for Healthcare (PervasiveHealth) and Workshops*, pages 139–146. IEEE, 2012.

[26] C. C. Hsia, K. Liou, A. Aung, V. Foo, W. Huang, and J. Biswas. Analysis and comparison of sleeping posture classification methods using pressure sensitive bed system. In *IEEE Int'l Conf. on Engineering in Medicine and Biology Society*, 2009.

[27] M.-K. Hu. Visual pattern recognition by moment invariants. *IEEE Trans. on Information Theory*, 8(2):179–187, 1962.

[28] W. Huang, A. A. P. Wai, S. F. Foo, J. Biswas, C.-C. Hsia, and K. Liou. Multimodal sleeping posture classification. In *IEEE Int'l Conf. on Pattern Recognition*, 2010.

[29] C. Idzikowski. Sleep position gives personality clue. *BBC News (September 16, 2003)*, 2003.

[30] Itseez. Open source computer vision library. `https://github.com/itseez/opencv`, 2015.

[31] M. J. Johnson and A. S. Willsky. Bayesian nonparametric hidden semi-markov models. *Journal of Machine Learning Research*, 14(Feb):673–701, 2013.

[32] S. Karanam, M. Gou, Z. Wu, A. Rates-Borras, O. Camps, and R. J. Radke. A systematic evaluation and benchmark for person re-identification: Features, metrics, and datasets. *arXiv preprint arXiv:1605.09653*, 2016.

[33] R. M. Khoury, L. Camacho-Lobato, P. O. Katz, M. A. Mohiuddin, and D. O. Castell. Influence of spontaneous sleep positions on nighttime recumbent reflux in patients with gastroesophageal reflux disease. *The American journal of gastroenterology*, 1999.

[34] I. Koprinska, G. Pfurtscheller, and D. Flotzinger. Sleep classification in infants by decision tree-based neural networks. *Artificial intelligence in Medicine*, 8(4):387–401, 1996.

[35] C.-H. Kuo, F.-C. Yang, M.-Y. Tsai, and L. Ming-Yih. Artificial neural networks based sleep motion recognition using night vision cameras. *Biomedical Engineering: Applications, Basis and Communications*, 16(02):79–86, 2004.

[36] C. H. Lampert, H. Nickisch, and S. Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *IEEE Proc. of Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2009.

[37] C. Lea, J. Facker, G. Hager, R. Taylor, and S. Saria. 3d sensing algorithms towards building an intelligent intensive care unit. *AMIA summits on translational science proceedings*, 2013:136, 2013.

[38] A. Lewicke, E. Sazonov, M. J. Corwin, M. Neuman, and S. Schuckers. Sleep versus wake classification from heart rate variability using computational intelligence: consideration of rejection in classification models. *IEEE Trans. on Biomedical Engineering*, 55(1):108–118, 2008.

[39] S. Liao, Y. Hu, X. Zhu, and S. Z. Li. Person re-identification by local maximal occurrence representation and metric learning. In *IEEE Proc. of Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[40] W.-H. Liao and C.-M. Yang. Video-based activity and movement pattern analysis in overnight sleep studies. In *IEEE Int'l Conf. on Pattern Recognition*, 2008.

[41] A.-A. Liu, Y.-T. Su, W.-Z. Nie, and M. Kankanhalli. Hierarchical clustering multi-task learning for joint human action grouping and recognition. *IEEE transactions on pattern analysis and machine intelligence*, 39(1):102–114, 2017.

[42] A.-A. Liu, N. Xu, W.-Z. Nie, Y.-T. Su, Y. Wong, and M. Kankanhalli. Benchmarking a multimodal and multiview and interactive dataset for human action recognition. *IEEE Transactions on cybernetics*, 2017.

[43] L. Liu, L. Shao, X. Li, and K. Lu. Learning spatio-temporal representations for action recognition: A genetic programming approach. *IEEE transactions on cybernetics*, 46(1):158–170, 2016.

[44] P. Liu, W. Liu, and H. Ma. Weighted sequence loss based spatial-temporal deep learning framework for human body orientation estimation. In *Multimedia and Expo (ICME), 2017 IEEE International Conference on*, pages 97–102. IEEE, 2017.

[45] W. Liu, Y. Zhang, S. Tang, J. Tang, R. Hong, and J. Li. Accurate estimation of human body orientation from rgb-d sensors. *IEEE Transactions on Cybernetics*, 43(5):1442–1452, 2013.

[46] B. D. Lucas and T. Kanade. An iterative image registration technique with an application to stereo vision. *Proceedings DARPA image Understanding*, page 121430, 1981.

[47] S. Maji, L. Bourdev, and J. Malik. Action recognition from a distributed representation of pose and appearance. In *IEEE Proc. of Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2011.

[48] N. Martinel, G. L. Foresti, and C. Micheloni. Person reidentification in a distributed camera network framework. *IEEE Transactions on Cybernetics*, 2016.

[49] S. Morong, B. Hermsen, and N. de Vries. Sleep position and pregnancy. In *Positional Therapy in Obstructive Sleep Apnea*. Springer, 2015.

[50] P. E. Morris. Moving our critically ill patients: mobility barriers and benefits. *Critical care clinics*, 2007.

[51] G. Nebehay and R. Pflugfelder. Clustering of Static-Adaptive correspondences for deformable object tracking. In *IEEE Proc. of Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[52] S. Obdržálek, G. Kurillo, J. Han, T. Abresch, R. Bajcsy, et al. Real-time human pose detection and tracking for tele-rehabilitation in virtual reality. *Studies in health technology and informatics*, 173:320–324, 2012.

[53] H. S. of Medicine. Finding Top-Line Opportunities in a Bottom-Line Healthcare Market. Technical report, Harvard School of Med., 2016.

[54] A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. In *Springer Int'l Journal of Computer Vision (IJCV)*, 2001.

[55] N. Padoy, D. Mateus, D. Weinland, M.-O. Berger, and N. Navab. Workflow monitoring based on 3d motion features. In *Computer Vision Workshops (ICCV Workshops), 2009 IEEE 12th International Conference on*, pages 585–592. IEEE, 2009.

[56] R. Panda and A. R. Chowdhury. Multi-view surveillance video summarization via joint embedding and sparse optimization. *IEEE Transactions on Multimedia*, 2017.

[57] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[58] X. Peng, L. Wang, X. Wang, and Y. Qiao. Bag of visual words and fusion methods for action recognition: Comprehensive study and good practice. *Computer Vision and Image Understanding*, 150:109–125, 2016.

[59] T. Penzel and R. Conradt. Computer based sleep recording and analysis. *Sleep medicine reviews*, 4(2):131–148, 2000.

[60] L. R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

[61] S. Ramagiri, R. Kavi, and V. Kulathumani. Real-time multi-view human action recognition using a wireless camera network. In *Int'l IEEE Conf. on Distributed Smart Cameras*, 2011.

[62] M. Raptis and L. Sigal. Poselet key-framing: A model for human activity recognition. In *IEEE Proc. of Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2013.

[63] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi. You only look once: Unified, real-time object detection. In *IEEE Proc. of Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[64] C. Sahlin, K. A. Franklin, H. Stenlund, and E. Lindberg. Sleep in women: normal values for sleep stages and position and the effect of age, obesity, sleep apnea, smoking, alcohol and hypertension. *Sleep medicine*, 2009.

[65] Y. Shi, Y. Tian, Y. Wang, and T. Huang. Sequential deep trajectory descriptor for action recognition with three-stream cnn. *IEEE Transactions on Multimedia*, 2017.

[66] J. Shotton, R. Girshick, A. Fitzgibbon, T. Sharp, M. Cook, M. Finocchio, R. Moore, P. Kohli, A. Criminisi, A. Kipman, et al. Efficient human pose estimation from single depth images. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 2013.

[67] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[68] L. Soban, S. Hempel, B. Ewing, J. N. Miles, and L. V. Rubenstein. Preventing pressure ulcers in hospitals. *Joint Commission Journal on Quality and Patient Safety*, 2011.

[69] J. Song, H. Jegou, C. Snoek, Q. Tian, and N. Sebe. Guest editorial: Large-scale multimedia data retrieval, classification, and understanding. *IEEE Transactions on Multimedia*, 2017.

[70] B. Soran, A. Farhadi, and L. Shapiro. Generating notifications for missing actions: Don't forget to turn the lights off! In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4669–4677, 2015.

[71] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–9, 2015.

[72] The Raspberry Pi Foundation. *Raspberry Pi 3 Model B*, 2017 (accessed July 17th, 2017). https://www.raspberrypi.org/products/raspberry-pi-3-model-b/.

[73] P. R. Torrens. The health care team members: Who are they and what do they do. *Collaboration across the disciplines in health care*, 2010.

[74] C. Torres, A. J. Bency, J. C. Fried, and B. S. Manjunath. Ram: Role representation and identification from combined appearance and activity maps. In *ACM/IEEE Proc. of Int'l Conf. on Distributed Smart Cameras (ICDSC)*, 2017.

[75] C. Torres, V. Fragoso, S. D. Hammond, J. C. Fried, and B. S. Manjunath. Eye-cu: Sleep pose classification for healthcare using multimodal multiview data. In *Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2016.

[76] C. Torres, J. C. Fried, K. Rose, and B. Manjunath. Deep eye-cu (decu): Summarization of patient motion in the icu. In *European Conference on Computer Vision (ECCV)*. Springer, 2016.

[77] C. Torres, S. D. Hammond, J. C. Fried, and B. S. Manjunath. Multimodal pose recognition in an icu using multimodal data and environmental feedback. In *International Conference on Computer Vision Systems (ICVS)*. Springer, 2015.

[78] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri. Learning spatiotemporal features with 3d convolutional networks. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 4489–4497. IEEE, 2015.

[79] V. Veeriah, N. Zhuang, and G.-J. Qi. Differential recurrent neural networks for action recognition. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4041–4049, 2015.

[80] R. Vezzani, D. Baltieri, and R. Cucchiara. People reidentification in surveillance and forensics: A survey. *ACM Computing Surveys (CSUR)*, 2013.

[81] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. Van Gool. Temporal segment networks: towards good practices for deep action recognition. In *European Conference on Computer Vision*, pages 20–36. Springer, 2016.

[82] T. Wang, S. Gong, X. Zhu, and S. Wang. Person re-identification by video ranking. In *Proc. of European Conf. on Computer Vision (ECCV)*, 2014.

[83] G. L. Weinhouse and R. J. Schwab. Sleep in the critically ill patient. *Sleep-New York Then Westchester-*, 29(5):707, 2006.

[84] C. Wu, A. H. Khalili, and H. Aghajan. Multiview activity recognition in smart homes with spatio-temporal features. In *Proceedings of the Fourth ACM/IEEE International Conference on Distributed Smart Cameras*, pages 142–149. ACM, 2010.

[85] J. Wu, Y. Zhang, and W. Lin. Good practices for learning to recognize actions using fv and vlad. *IEEE transactions on cybernetics*, 46(12):2978–2990, 2016.

[86] Y. Yang, C. Deng, S. Gao, W. Liu, D. Tao, and X. Gao. Discriminative multi-instance multitask learning for 3d action recognition. *IEEE Transactions on Multimedia*, 2017.

[87] Y. Yang and D. Ramanan. Articulated human detection with flexible mixtures of parts. *IEEE Trans. on Pattern Analysis and Machine Intelligence,*, 2013.

[88] H. Yu, M. Li, H.-J. Zhang, and J. Feng. Color texture moments for content-based image retrieval. In *IEEE Proc. of Int'l Conf. on Image Processing (ICIP)*, 2002.

[89] R. Zhao, W. Ouyang, and X. Wang. Person re-identification by salience matching. In *IEEE Proc. of Int'l Conf. on Computer Vision (ICCV)*, 2013.

[90] R. Zhao, W. Ouyang, and X. Wang. Unsupervised salience learning for person re-identification. In *IEEE Proc. of Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2013.

[91] R. Zhao, W. Ouyang, and X. Wang. Learning mid-level filters for person re-identification. In *IEEE Proc. of Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[92] X. Zhen, F. Zheng, L. Shao, X. Cao, and D. Xu. Supervised local descriptor learning for human action recognition. *IEEE Transactions on Multimedia*, 2017.